Face recognition based on image sets

Huang, Likun

2014

Huang, L. (2014). Face recognition based on image sets. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/61822

https://doi.org/10.32657/10356/61822

Face Recognition based on Image Sets

Huang Likun

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2014

Acknowledgments

I have many people to thank for their help and support during the development of this thesis. To me, they have been exemplary personal and professional mentors. Foremost among them is Prof. Tan Yap-Peng. As my Ph.D supervisor, Professor Tan always has good ideas to share, and provides easy ways to begin with. He has always been energetic and full of passion. Whenever I meet with problems in my research or daily life, he is there ready to provide valuable comments and suggestions. He taught me how to read and write in English, how to formulate research problems and solve them, everything from zero with great patience and expertise. I would like to express my deepest gratitude and appreciation to Prof. Tan for what he has done for me. His trust and support have been the main source of inspiration and impetus for me.

The other important person I would like to thank is Dr. Lu Jiwen. Throughout the past few years, he has given me valuable lessons covering pattern recognition, face recognition, machine learning, etc. Besides, he has always provided valuable comments from some different perspectives when I made progress. I have learnt a lot from him about how to improve the finer points of technical writing and how to come up with a decent response letter to address the reviewers' questions. This thesis would not be in its current form without his constant guidance and constructive comments.

I would also like to thank the wonderful study group that discussed and exchanges advances in pattern recognition, face recognition, machine learning and other related topics. They are Yang Gao, Hung Tzu-Yi, Li Maodong, Chuah Seong-Ping, and Liu

Nini. Without the meaningful discussions with them, I cannot have a comprehensive understanding of the related fields. Their in-depth knowledge in these fields and constructive comments also teach me a lot on how to carry out my research work. There are also many individuals who have directly or indirectly contributed to this thesis. I am thankful to Xiao Yong, Wang Junyan, Wang Tao, Yang Wei-Ting, Luo Shan, Luo Ye, and Wang Wenwen, for their friendship and encouragement in the past years. I would like also to thank all the anonymous reviewers who gave generously of their time and expertise.

Last but not least, I wish to dedicate this thesis to my family and thank for their unconditional love and understanding through out all these years of my study away from home.

Huang Likun

Singapore, July 2013

Contents

Abstract

In this thesis, we study the problem of *face recognition based on image sets*. The main objective of our work is to develop set-based distance metrics that are able to measure the similarity between image sets, rather than conventional distance metrics that can only measure the distance between samples.

The face images obtained from real-life impose great challenges to the conventional face recognition systems. Large variations in appearances and various imperfections such as occlusions and misalignments in the face images severely degrade the recognition performance. One possible solution is to utilize more face images of each person for testing, e.g., a collection of photos from personal galleries or frames extracted from a video clip. Under such circumstances, the face recognition task becomes the process of modeling and matching image sets. Our investigation then focuses on developing appropriate models and set-based distance metrics for representing and matching different image sets.

In general, the methods for solving the problem of face recognition based on image sets can be divided into two predominant categories, namely the single model-based methods and the multi-model based methods. Firstly, we investigate the single model-based methods to exploit its nice properties such as computational efficiency. Through a close look into the existing methods, we propose a generalized subspace distance (GSD) framework to illustrate the underlying relationships among the existing methods, which can be considered as special cases of the proposed framework. In view of the unsupervised

matching adopted in the existing single model-based methods, we introduce a parameter, which is learned from a discriminative learning procedure, into the proposed subspace distances, namely fractional order weighted subspace distance (FOWSD) and exponential weighted subspace distance (EWSD). Furthermore, we extend the proposed FOWSD and EWSD to their affine versions such that each image set can be better represented.

Subsequently we proceed to investigate the multi-model based methods, which have stronger capabilities to describe image sets with more complex distributions than the single model-based methods. Generally, the multi-model based methods consist of two steps: the local model construction and the manifold-to-manifold distance measure. We first propose a novel co-learned multi-view spectral clustering (CMSC) approach, which integrates information from multiple views of the same image set to enhance the performance of the local model construction. Then we propose a new manifold-to-manifold distance measure (CRMMD) that computes the distance between the local model and its approximation in a collaborative manner, which is more robust to the distortions caused by outliers.

Similarly, the multi-model based methods that we investigated are based on unsupervised matching. To exploit the advantage of supervised learning where the discriminative information provided by class labels are fully utilized in the training phase, we propose a novel multi-manifold metric learning method (MMML) that adds a discriminative learning stage before proceeding to image set matching. Different from the existing supervised learning approaches that learn only one subspace for all the image sets, the proposed MMML method learns a collection of person-specific distance metrics, one for each class, to match manifolds in different classes, such that the recognition performance can be significantly improved.

To verify the effectiveness of all the proposed methods, we have conducted extensive experiments using several popular face databases.

List of Figures

List of Tables

Chapter 1

Introduction

Our research work in this thesis focuses on face recognition based on image sets (FRBIS). In this chapter, we first discuss the motivation of our research in a holistic and systematic manner, where our objectives are established. Then the main contributions of our work will be summarized.

1.1 Motivations and Objectives

- **Why automatic face recognition? — Human ability to distinguish different faces is not perfect.**

  Adults have the ability of recognizing the face of a special person from thousands of other faces from different individuals even under challenging conditions, e.g., human faces look alike compared with other general objects and there are usually large variations in the expression, pose, illumination and occlusion that affect the appearance of the same person. However, such attractive ability is associated with some limitations, e.g.,

  (i) The robust abilities of adults require not only specialized perceptual processes and neural mechanisms [1], but also accumulated experiences during many years (from infancy to adolescence) [2] [3].

  (ii) The limitation of the human brain is that the human brain can only remember the faces of limited number of individuals accurately and is not able to remember every detail in a face.

To break the limitations described above, researchers attempted to imitate the excellent face recognition ability of humans by utilizing computer system that has such characteristics as powerful computational capability and large storage capacity. In the 1970's, the earliest work on automatic face recognition [4] [5] appeared and succeeded. Since then, the automatic face recognition techniques exhibit great potentials in a wide range covering commercial and security applications. Over the past few decades, automatic face recognition has attracted great research interest and grown quickly across many disciplines such as artificial intelligence, computer vision, pattern recognition, image processing, neural network, etc. A large number of classical methods of face recognition based on single images have been proposed in the literature [6] [7] [8] [9] [10] [11], which have significantly improved the performance of face recognition systems under controlled conditions. The controlled condition means that the variables that affect the acquirement of visual information of faces are fixed or constrained in a small range, e.g., indoor environment with satisfied lighting (neither too weak nor too strong) or slight variations in pose and expression of an individual.

- **Why face recognition based on image sets? — Challenging conditions require more information.**

  With the rapidly developing Internet and constantly increasing online social interactions, a large number of social media sharing websites have emerged, such as Facebook, YouTube, Flickr, etc. Everyday hundreds of millions of people are posting their personal photos and videos to these social media networks. With massive data, video sequences and multiple still images of an individual are much easier to be obtained than before. However, these personal images including faces are usually acquired under unconstrained conditions where there are large variations in illumination, expression, pose, resolution, etc. In such practical environments, it is still very difficult for the conventional face recognition approaches to provide a reliable and acceptable recognition performance. In the conventional face recognition scenario shown in Fig. 1.1, given a gallery set where each reference is a single-shot face image, a probe (or query) is also a single-shot image and the similarity between a gallery and a probe is obtained by calculating the image-to-image distance.

To enhance the performance of conventional face recognition approaches under unconstrained conditions, a possible solution is to exploit more face images for each

Figure 1.1: Conventional face recognition techniques recognize a person from a single-shot image.

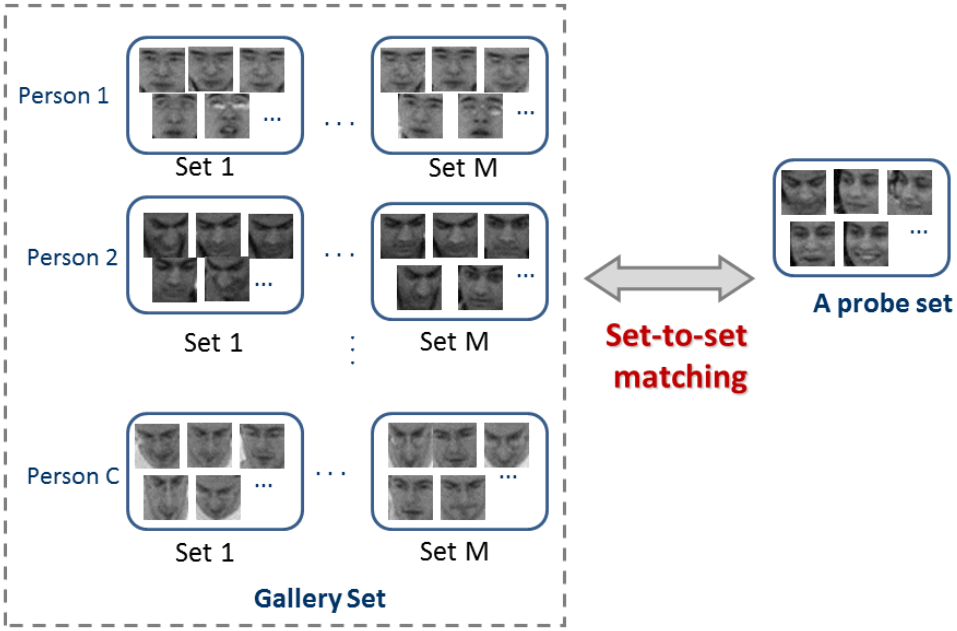

Figure 1.2: Scenario of face recognition based on image sets.

person in classification tasks. Compared with a single-shot face image, an image set of a person is able to provide more complete information that covers variations in the individual's appearance. Generally, an image set could comprise multiple still images that are captured under various circumstances, or frames cropped from a video clip (sometimes these frames are not in continuous sequence). These image sets can be easily acquired through various social media networks as mentioned above. Since more discriminative information can be extracted from an image set and utilized to model the appearance of an individual's face, the face recognition approaches based on image sets have the potential of achieving a better face recognition performance than the conventional methods. Inspired by this observation, our investigations in this thesis mainly focus on the topic of face recognition based on image sets, which recognizes an individual by utilizing a set of his/her face images. Fig. 1.2 illustrates a typical scenario of face recognition based on image sets. Given a gallery set where each reference is an image set, a probe (or query) is also an image set and the similarity between a gallery and a probe is obtained by calculating the set-to-set distance.

- **Objectives**

  The main objectives of our research work in this thesis are to better understand the problem of face recognition based on image sets in a comprehensive and systematic manner, and to develop the set-based distance metrics that are able to measure the similarity between image sets rather than the conventional distance metrics that can only measure the distance between single-shot samples. The scope of this thesis covers topics from the single-model based methods to the multi-model based methods, from the single-view based clustering methods to the methods based on multi-view data, from the unsupervised methods to the supervised-learning methods, etc. To achieve these objectives, the issues that need to be addressed are summarized in the following:

  (i) Analyze and better understand the characteristics of the problem of face recognition based on image sets that aims to enhance the performance of conventional face recognition approaches under challenging conditions.

(ii) Investigate and propose novel methods to enhance the performance of face recognition based on image sets, which includes: the single-model based methods, multi-model based methods, unsupervised set-based matching methods, supervised learning methods and multi-view based methods.

(iii) Evaluate the performance of the state of the arts and demonstrate the effectiveness of the proposed methods through extensive empirical comparisons.

1.2 Main Contributions

In this thesis, we focus on the problem of face recognition based on image sets. Relating to this problem, there are issues need to be addressed from two aspects: (1) the model that is utilized to describe an image set and (2) the distance metric that measures the similarity between two such models.

In terms of the models, we group the existing approaches into three categories: the model-free methods, the single-model based methods and the multi-model based methods, among which the single-model based methods and the multi-model based methods are predominant. We will start our research from the single model-based methods that will be discussed in Chapter 3. Although our proposed methods benefit from the efficient property of the single model-based methods, it will be observed that the single model has limited ability to capture the underlying structure of a data set with complex distribution [12]. Thus, a more appropriate way to describe such an image set is to utilize multiple models. As shown in Fig. 1.3, our investigation moves from the single model-based methods in Chapter 3 to the multi-model based methods in Chapter 4, 5 and 6.

Towards the issue of designing the distance metrics, we study the unsupervised methods for matching face image sets in Chapters 4 and 5. It is well known that the unsupervised matching methods are computationally efficient but not discriminative enough for classification. Thus, we further investigate the supervised-learning methods for matching face image sets in Chapter 6.

The whole framework of this thesis is illustrated in Fig.1.3, which demonstrates the progressive relationships among the four pieces of work investigated in Chapter 3, Chapter 4, Chapter 5 and Chapter 6. Based on the above framework, we further summarize the contributions of our work in this thesis by answering the questions below:

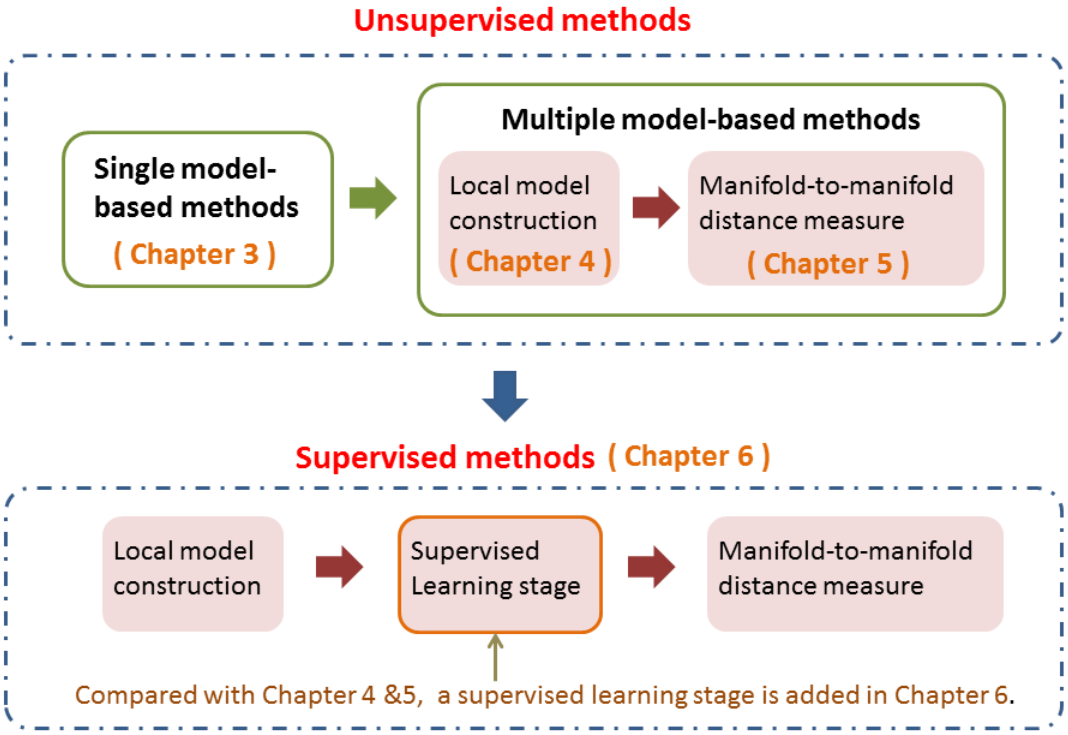

Figure 1.3: The whole framework of this thesis.

- ***For the single-model based methods, is it possible to discriminatively measure the similarity between two image sets?***

  Among the single-model based methods, a predominant subcategory is the correlation -based method that models an image set by using linear representations such that these methods are usually computationally efficient and robust to outliers. There are various linear representations proposed in the literature, e.g., linear subspaces, affine representations and convex hulls. However, some limitations appear in existing methods: (1) Most existing methods for measuring subspace distances ignore the label information of training image sets; (2) the conventional distances between convex hulls or affine hulls require high computational complexity. To solve the above problems, a novel distance metric called generalized subspace distance (GSD) framework is proposed to provide an explicit rule for measuring distances between subspaces. Furthermore, we make the GSD more discriminative by adding some pre-learned coefficients to improve the recognition rate and extend

the proposed subspace distances to their affine versions. The efficacy of the proposed method will be shown later in Chapter 3 through extensive evaluations and comparisons.

- ***For the multi-model based methods, how to construct more representative local models?***

  The multi-model based methods consists of two stages: (1) local model construction and (2) manifold-to-manifold distance measure. We investigate the first stage in Chapter 4 and the second stage in Chapter 5. To construct a local model, a clustering method such as K-means algorithm should be adopted. However, the current clustering methods only utilize the intensity values of an image set. Actually, multiple views of an image set can be easily obtained, e.g., we can extract the local binary pattern (LBP) [13] [14] descriptors or the scale-invariant feature transform (SIFT) [15] descriptors from a given image set. Based on this observation, we propose a Co-learned Multi-view Spectral Clustering (CMSC) method in Chapter 4 that combines more types of information to enhance the clustering performance such that a more representative cluster and its corresponding local model can be obtained. The improved performance of the proposed CMSC method will be demonstrated through experiments.

- ***For the multi-model based methods, is there a more robust measurement for calculating the manifold-to-manifold distance?***

  In the multi-model based methods, an image set with complex distribution is often considered as a nonlinear manifold. To describe this manifold, an image set is often divided into several subsets, each of which is described using a local model such that a manifold can be represented by a collection of local models. Since the multi-model based methods just emerged recently, the problem of how to measure the similarity between two manifolds has received less attention. An existing manifold-to-manifold distance (MMD) was proposed by Wang et al. [16]. In their method, the pairwise distances among all the local models are calculated from which the minimum value is selected as the manifold-to-manifold distance. However, computing MMD in this way may lead to unstable recognition performance due to the fact

that MMD method is usually sensitive to outliers. To solve this problem, we propose a collaborative Reconstruction-based Manifold-Manifold Distance (CRMMD) method that calculates the distance between a local model and its collaborative approximation, thus avoiding the unstable recognition performance. The efficacy and robustness of the proposed method will be illustrated later in Chapter 5 through experiments.

- ***For the multi-model based methods, is it possible to further strengthen the discriminability of the existing supervised learning methods for measuring the similarity between image sets?***

  Upon investigating the unsupervised methods for matching image sets in Chapters 4 and 5, we naturally proceed to the supervised learning methods that explore more discriminative information of the training datasets for enhancing the performance of image set matching. It is observed that the existing discriminant analysis approaches utilize label information and seek only one common space for all the training data from different persons. In contrast, we propose a discriminative learning method called Multi-Manifold Metric Learning method (MMML) that adopts a person-specific manner to learn a collection of distance metrics, one for each class. The improved performance of the proposed method will be demonstrated in Chapter 6.

1.3 Thesis Organization

The rest of this thesis is organized as follows.

In Chapter 2, the problem formulation, structure of the system, possible applications and challenges on automatic face recognition techniques are introduced, and the research achievements in this area are summarized. Specifically, the related works on the problem of face recognition based on image sets (FRBIS) are reviewed.

Chapter 3 first reviews several popular single-model based methods. Motivated by these methods, we propose a Generalized Subspace Distance (GSD) framework and show that most existing subspace similarity measures can be considered as its special cases. Within the GSD framework, we further propose a Fractional Order Weighted Subspace

Distance (FOWSD) method and an Exponential Weighted Subspace Distance (EWSD). Furthermore, the proposed FOWSD and EWSD methods are extended to their affine versions. Experimental results on two set-based classification tasks, including face recognition/object recognition based on image sets, are presented to show the effectiveness of the proposed method.

Motivated by the discussion about single-model based methods, we proceed to investigate the multi-model based methods in Chapters 4 and 5. To obtain a more representative local model for each subset, we propose a Co-learned Multi-view Spectral Clustering (CMSC) approach in Chapter 4 to fully exploit the multiple types of information for an image set. By modeling a manifold (an image set) as a collection of local models, we first discuss the existing achievements on the manifold-to-manifold distance in Chapter 5. Then we propose a Collaborative Reconstruction-based Manifold-Manifold Distance (CRMMD) method to handle some challenging conditions where the image sets contain severe noises. Experimental results are illustrated to show the efficiency of the proposed method.

Based on the unsupervised method proposed in Chapters 4 and 5, we proceed to investigate the supervised learning methods in Chapter 6. To enhance the discriminability of the supervised learning methods for measuring the similarity between manifolds, a Multi-Manifold Metric Learning method (MMML) is proposed. Our method is extensively evaluated on three popular face databases and compared to the state of the arts.

Chapter 7 concludes this thesis. Several directions of the future work are also discussed.

Chapter 2

Background and Literature Review

2.1 Automatic Face Recognition

2.1.1 Problem Statement

- **Why use face for recognition?**

  Before biometric-based techniques appeared, a person usually visits an area with protected information by utilizing a password, key, token, national ID, driver license, passport or smart card. However, the key, national ID, driver license, passport or smart card might be stolen or damaged, and the password might be forgotten. Compared with these physical passes, a more reliable way to access privacy is to utilize the biometric-based identity recognition techniques that determine a specific person through his/her physiological characteristics or behavioral patterns such as face, fingerprint, iris, palm, ear, voice and gait. These biological characteristics of a person are unique, irreplaceable and available all the time. Among the biometric-based identification techniques, face recognition shows more potential merits over others. That is, most biometric-based methods such as iris, fingerprint, palm and ear, need the cooperation of users, putting their thumbprints or scanning the irises. In contrast, there is no strict collaboration requirement for face recognition systems where face images can be captured by public video surveillance or even taken by hidden cameras at a distance. Furthermore, iris images and voices are usually collected through special or expensive equipments or under specific conditions such as a place without much noise. In contrast, face images can be easily

obtained from any video surveillance or camera at a reasonable price. Compared with other biometric-based techniques, the inherent advantages of face recognition have made an profound impact on various applications.

- **Problem Statement of face recognition**

  The problem of face recognition through visual information of faces can be formulated as follows: given a still image/a video clip that contains human faces and a gallery of face database available for comparison, how to verify or determine the identities of persons through their faces cropped from the image or video clip?

- **Applications**

  Over the past several decades, face recognition has attracted great research interest and become a hot research topic in the field of visual recognition due to its vast commercial values and security applications.

  (i) **Information security**: For security purposes, a captured frontal face photo is usually required by the access control to checkpoint, buildings, even computers and smart handphone logon [17] [18] [19].

  (ii) **Smart cards**: Photographs on national IDs, driver's licenses, passports, visa cards, students' matriculation cards.

  (iii) **Multi-media applications**: Video games such as kinect, training programs, monitoring patients who need to do rehabilitative exercises in health center, monitoring the behaviors of babies or senior people whose self-care abilities are poor, monitoring and analyzing customers' preferences [20] [21].

  (iv) **Law enforcement and surveillance**: Video surveillance is a very important way to search criminals and suspects for maintaining social stability and protecting people's lives and properties [22].

  (v) **Other applications**: Searching missing persons through mugshot albums or videos, forensics, witness face reconstruction, post-event analysis, etc. Recently, face recognition techniques have developed rapidly and been applied

into many related areas such as gender recognition/classification, age estimation, emotion recognition, local feature location/detection (e.g. patches of eyes, nose and mouth) [23] [24] [25] [26] [27] [28] [29] [30].

The above applications of face recognition cover various research fields including pattern recognition, computer vision, image processing, video indexing, analysis, etc. The achievements in face recognition techniques will promote the development of many other related subjects and benefit from these disciplines at the same time.

- **Challenges**

  Face recognition is a special case of pattern recognition problems. However, since most of the face images (such as frontal view) of different persons are alike to each other in the appearances, the conventional pattern recognition techniques, e.g. object recognition techniques, cannot be directly performed to discriminate a person through his/her face images [31]. Moreover, the appearances of faces change easily due to many factors. In [32] [33], these factors are divided into two categories: intrinsic factors and extrinsic factors. In the category of intrinsic factors, the geometrical structure of a 3D face changes due to facial expression and aging of the same person, and that varies between different persons. In the category of extrinsic factors, the appearance of a face is sensitive to the variations in illumination, imaging, noise, pose (e.g. scarf, hair style and make-up), resolution, misalignment, etc. Through a thorough review of the state of the arts, e.g., FERET evaluations [34] [35], FRVT2000 [36], FRVT2002 [37], FAT2004 [38], FRGC [39] and FRVT2006 [40], etc., the following issues are suggested as the major difficulties in the field of face recognition by most researchers.

  (i) **Illumination variations**: In many practical applications, the problem of illumination-invariant face recognition is still challenging and difficult under complex lighting environments. In a face recognition system, it may be difficult to obtain the accurate visual information of faces under an unsatisfied direction of lighting source [41] [42]. Furthermore, a face usually appears to be very different under varying lighting conditions, which results in the larger

intra-person appearance variations than the inter-person appearance variations. The advantages of illumination-invariant face recognition techniques have potentials in many applications, such as outdoor environments under strong sunlight or under low lighting at night.

(ii) **Pose variations**: Recognizing a face with an arbitrary rotation is another challenging issue confronted by researchers. For a conventional face recognition system, it is a challenging task to extract the intrinsic characteristics that are independent on the varying poses from faces with large appearance variations due to varying view-points and 3D-2D transformations. The pose-invariant face recognition techniques aim to solve the problem of recognizing uncooperative individuals under unconstrained conditions in real life [43].

(iii) **Misalignment**: In order to obtain a good classification performance, it is a critical step to align faces to a unified template according to their facial features [44] [45] [46]. However, the existing work on facial feature detection are still struggling in producing accurate locations of such features. Thus, the performance of most face recognition techniques still suffer from the problems of misalignment. The face recognition techniques for misalignment have the ability of dealing with face images obtained from unconstrained conditions where faces are not easy to be aligned due to variations in pose, occlusion, etc.

2.1.2 System Structure

Fig. 2.1 illustrates the flowchart of a face recognition system that contains four key modules: (1) face detection, (2) face alignment/normalization, (3) face representation (including feature extraction and learning procedures), and (4) face classification/identification. Sometimes the functions of these four modules are not totally independent, e.g., face detection and facial feature detections (i.e. eyes, mouth and nose) can be achieved simultaneously. Next, we describe the respective functions of these four modules in detail.

- **Face Detection**

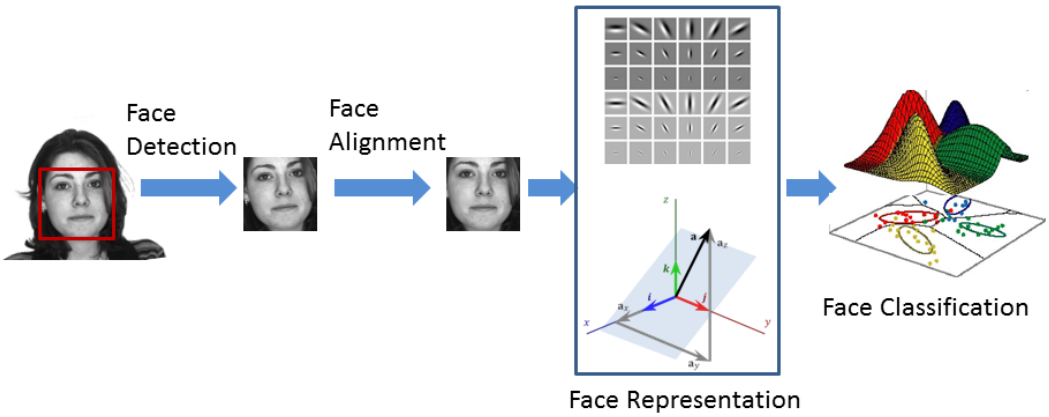

Figure 2.1: The flowchart of a face recognition system.

Given an input image, it needs to be determined whether there exists any face. If a face does exist, the position of this face and its size need to be detected.

Having evolved over the past few decades, the face detection techniques become the most mature component among the four modules of a face recognition system. Numerous algorithms have been developed and applied to various practical application scenarios such as the face detection and tracking functions of a digital camera. The most popular face detection technique is the Viola-Jones algorithm [47], which concatenates a series of weak classifiers to construct a strong cascaded classifier. Based on this classic approach, many improved algorithms have been proposed to enhance the performance of face detection recently.

- **Face Alignment**

  The performance of most face recognition algorithms rely on the module of face alignment, i.e., accurately transforming faces into the same canonical template through a series operations such as scaling, rotation, segmentation, etc. Generally, face alignment requires known positions of important facial features, e.g. eyes, nose and mouth, which can be manually labeled or learned through a class-specified learning procedure.

  The challenges of face alignment techniques exist in using a unified mathematic model to match various types of faces. To achieve this goal, numerous algorithms

have emerged recently. Among them, some typical methods are widely recognized to be of high qualities and able to provide good performance, such as Active Shape Model (ASM) [48], Active Appearance Model (AAM) [49], 3D Morphable Model [50], etc. For a better illustration, some exemplar face images before and after alignment are demonstrated in Fig. 2.2.

- **Face Representation**

  Before proceeding to the classification step, a face needs to be well represented. Generally, there are two different ways to represent a face: (1) Pre-define a template for calculating the descriptors of a face, e.g. the Local Binary Pattern (LBP) [14] feature and the Scale-Invariant Feature Transform (SIFT) [15] feature, as shown in Fig. 2.3; (2) According to a certain criterion, parameters are learned from training faces for optimally representing both gallery and probe faces, e.g. the Principal Canonical Analysis (PCA) [6] method and Linear Discriminant Analysis (LDA) [7] method, as shown in Fig. 2.4.

- **Face Classification**

  upon the completion of the above steps, both the probe face image and the gallery of face database are represented by utilizing the parameters learned from the training faces. Then comparisons between these representations are conducted to determine

Figure 2.2: Some exemplar face images before and after the face alignment.

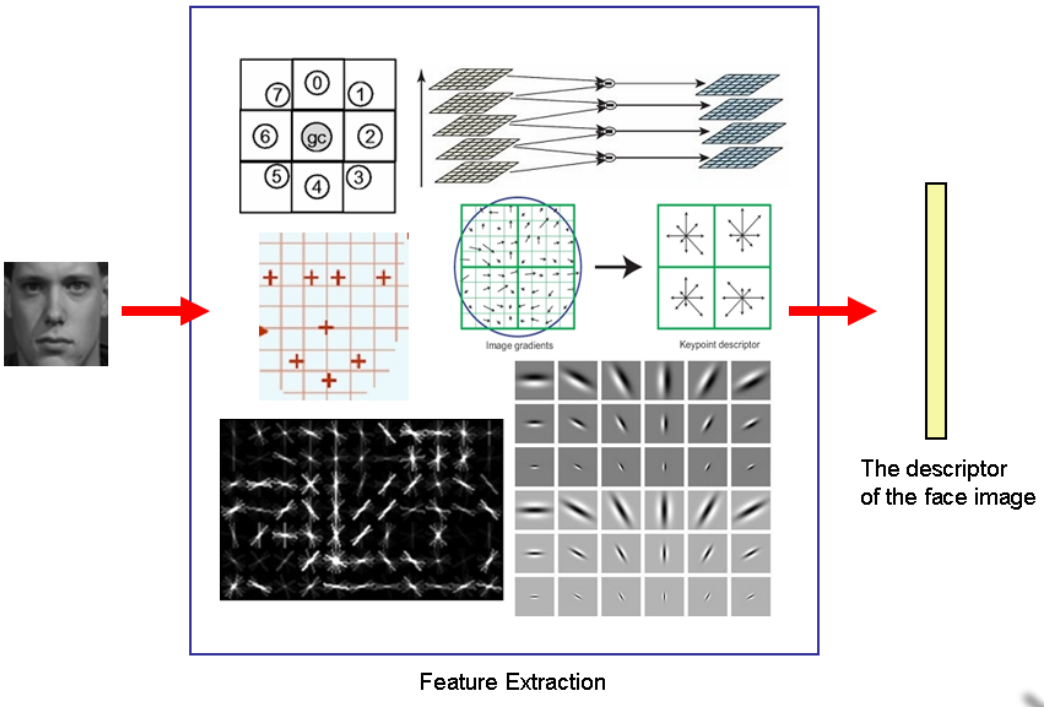

Figure 2.3: Calculating the descriptor of a face image through pre-defined templates.

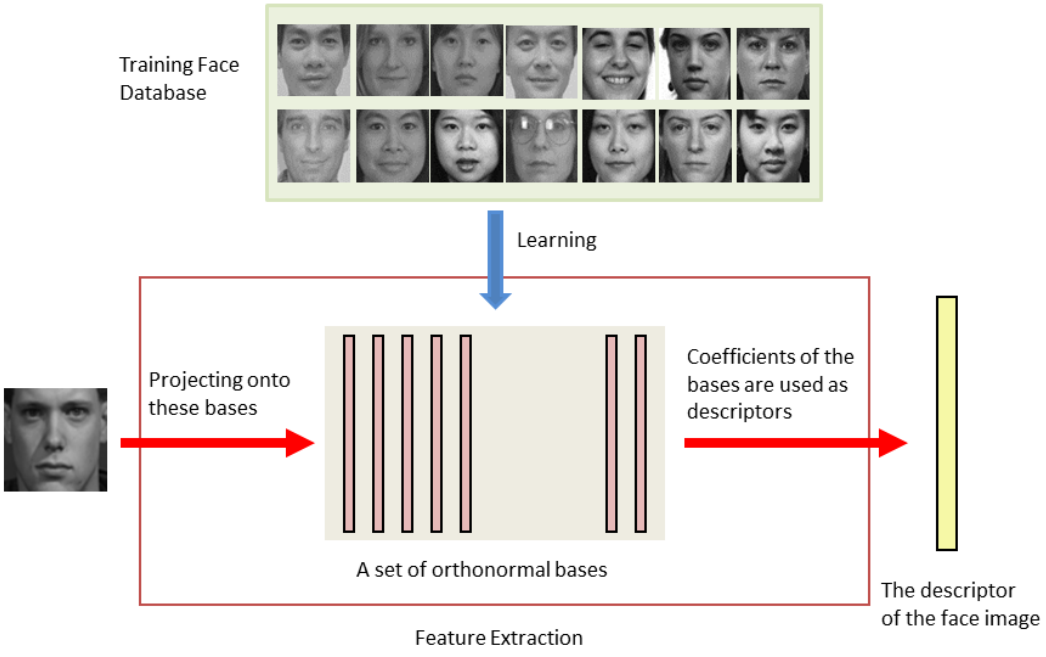

Figure 2.4: Calculating the descriptor of a face image through learned parameters.

the identities of individuals. From the above analysis, the face classification module is closely related to the face representation module, and the approaches adopted for face classification vary according to different representations of faces.

2.2 Literature Reviews

For a face recognition system, the face representation module is the most important component that attracts great research interest in recent years. A large number of achievements in face recognition techniques mainly contribute to the representation of faces. In this section, we divide the face recognition approaches into two categories: (1) feature-based methods and (2) appearance-based holistic methods. Some classic methods will be extensively evaluated.

2.2.1 Feature-Based Approaches

Given an input face image, the feature-based methods first detect the salient facial features, e.g. eyes, nose, mouth and head outline, and then extract the geometric information and topological structures of these facial features, e.g. the positions of eyes, the width of head, the angles of corners, the distances and ratios among these feature points, etc. Then these parameters are concatenated to constitute a low-dimensional feature vector to describe the high-dimensional input face images. Most of the feature-based approaches have been proposed before the year 2000.

- **Feature-based face recognition techniques in the 1970s**

  The earliest work on automatic face recognition appeared in the early 1970s [51] [4]. Generally, the early work utilized the measurements of facial attributes as facial descriptors that can be classified by using general pattern classification techniques. In Kanade's method [51], 16 measurements, such as the distances between corners and the ratios of distances, etc., were extracted to describe a face image and the Euclidean distance was utilized to measure the similarity between two face images. A small face database containing only 20 persons with 2 face images per person was used to evaluate the performance and a recognition rate of 75% was achieved on this

database. Similarly, in the method proposed by Kelly [4], the distances between important facial feature points were employed to represent each face image.

- **Feature-based face recognition techniques in the 1980s**

  During the 1980s, the development of face recognition techniques confronted a dormant period. There was not much improvement achieved in the performance of face recognition systems.

- **Feature-based face recognition techniques from the 1990s till now**

  During this period, various techniques were used in the feature-based face recognition systems to extract facial descriptors, such as Hough transform [52], deformable templates [53] [54] [55], etc. However, these algorithms share some common drawbacks as following: (1) They were heavily dependent on the heuristic conceptions; (2) It was difficult for these methods to adjust their parameters for conforming a facial structure perfectly; (3) Many of these algorithms detected and extracted facial features manually, e.g., a mixture-distance based method proposed by Cox et al. [56]. If these algorithms performed fully automatically, the recognition rates might drop significantly due to the low accuracies of the facial feature extraction techniques.

  To overcome the above drawbacks, methods based on the Hidden Markov Model (HMM) have been proposed to improve the performance of face recognition systems [57] [58] [59] [60]. In these methods, a face image was partitioned into multiple local patches, each of which was assigned with a state of the HMM. Then a probe face image was represented by a one-dimensional observation sequence and compared with each HMM in the gallery face database. Thus the best match with the highest likelihood was confirmed as the identity of the probe face. Furthermore, by utilizing the Principal Component Analysis (PCA) to extract the features of each region instead of directly using the intensity values, the method proposed by Nefian and Hayes [61] enhanced the recognition performance of HMM-based methods.

  A popular face recognition algorithm is the Elastic Bunch Graph Matching (EBGM) [62] [63] that is based on a well-known technique called Dynamic Link Architecture

(DLA) [64] [65]. Given the fiducial points of a face image, the EBGM method constructed a graph for this face based on these fiducial points, where each node in the graph was described by a jet that consists of a set of Gabor wavelet coefficients extracted from a window area centered at a fiducial point. The value of each arc is the distance between the corresponding fiducial points. Thus, a face image was described by a full connected graph. Furthermore, a set of jets, i.e., a face bunch graph, was attached to each node instead of a single jet. In the FERET evaluation [66] [34], the EBGM algorithm was one of the best performing methods in terms of the recognition rate due to imitating human visual system [67]. This motivated the EBGM-based face recognition systems to be applied to various practical scenarios, including face detection, gender recognition, pose estimation, etc.

In the feature-based face recognition techniques, another typical method was proposed in [68], which extended the self-organizing map (SOM) method for learning and utilized the convolutional neural network (CNN) method for robust feature extraction.

From the above analysis, there are some advantages of the feature-based approaches: (1) These algorithms show robust performance when face images contain large variations of misalignment, e.g. scaling, rotation and translation, which are common conditions in our real lives; (2) Instead of using all the pixels of a face image, it is much more concise and compact to represent a face by using features only from several important fiducial points; (3) Accordingly, the computational complexity of matching faces decreases significantly.

On the other hand, the feature-based approaches heavily depend on the performance of facial feature detection techniques. And the automatic and accurate detection of important points is still a difficulty. Moreover, the recognition performance might degrade when the detected fiducial points are less discriminative to represent a face.

2.2.2 Appearance-Based Holistic Approaches

In [69], Brunelli and Poggio showed their experimental results and concluded that the template-matching based methods outperform the feature-based matching approaches.

Although this conclusion stopped numerous researchers to plunge into the area of feature-based face recognition, it did start many new opportunities for the research of appearance-based face recognition techniques.

In the appearance-based holistic approaches, people attempted to describe a face image in a holistic manner, e.g., utilizing all the pixel values of a face rather than using several important fiducial points or local features. Generally, the appearance-based holistic approaches convert a two-dimensional face image into a one-dimensional vector (a sample point) in a high-dimensional space. In such a high-dimensional space sample points show sparsity due to the large volume of the space; However, the computations among vectors become a difficulty. This motivated the rapid development of dimension reduction approaches that have a positive impact on removing the curse of dimensionality.

Our discussions on the appearance-based holistic approaches can be chronologically summarized as four stages: (1) learning linear transformations, (2) learning nonlinear transformations, (3) using linear transformations to approximate nonlinear transformations and (4) new trends.

- **Learning linear transformations—PCA, LDA, ICA, etc.**

  PCA: The first attempt to use Principal Component Analysis (PCA) [70] [71] to compactly describe and reconstruct a face image was proposed in the references [72] [73]. In their work, it was indicated that the reconstruction of a face image could be approximated by linearly combing a small set of coefficients and the corresponding principal components. On this basis, a landmark called eigenfaces [74] [75] was proposed by Turk and Pentland to represent face images for detection and identification.

  PCA first calculates a set of linearly uncorrelated bases, i.e., the principal components, based on a set of sparse and correlated training data. These principal components are then used to compactly describe an image data, which can efficiently remove the statistical redundancies in the nature images, especially face images. Thus one advantage of using PCA to represent a face image is that the projection coefficients are much more insensitive to the noise than the original high-dimensional vector. Suppose that there is a training set of face images $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$

in a high-dimensional space $\mathbf{R}^D$, the objective of PCA is to seek a low-dimension linear subspace, i.e., a set of orthonormal bases, in which the sum of scatters of all the projected training data is maximized:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} (\mathbf{X}^T\mathbf{w})^T(\mathbf{X}^T\mathbf{w}) = \arg\max_{\mathbf{w}} \mathbf{w}^T(\mathbf{X}\mathbf{X}^T)\mathbf{w}$$
$$s.t. \qquad \mathbf{w}^T\mathbf{w} = 1, \tag{2.1}$$

where $\mathbf{w}^T\mathbf{w} = 1$ is only for a scaling factor and will not affect the optimization process.

As an extension of the eigenface method, Moghaddam and Pentland proposed a Bayesian method [76] that used the probability distance rather than the Euclidean distance to measure the similarities between images. The difficulty of this method lies in the probability estimation from limited training samples per person in such a high-dimensional space. To solve this problem, they designed an efficient approach to estimate the probability distribution by splitting the original space into two complementary spaces, i.e., a principal space and its orthogonal space.

ICA: The unsupervised statistical method PCA only explores the second-order moments of the face images, i.e., the pairwise relationships among pixels. However, it is possible that some important information lies in the high-order statistical relationships of face images, which could provide a better face recognition performance. In view of this, Independent Component Analysis (ICA) was proposed in references [77] [78] where the objective was to seek independent components rather than uncorrelated ones for representing face images.

LDA: Another well-known holistic method named Fisherface was proposed by Belhumeur et al. [79]. In this method, the authors observed that PCA could not effectively remove the variations caused by the changes of lighting conditions and facial expressions, such that the intra-class variations are often larger than the inter-class variations. Thus they utilized the Fisher's Linear Discriminant Analysis [80] to enhance the discriminability of the learned low-dimensional subspaces, where the ratio between the determinant of the inter-class scatter and that of the intra-class

scatter is maximized:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{|\mathbf{w}^T \mathbf{S}_B \mathbf{w}|}{|\mathbf{w}^T \mathbf{S}_W \mathbf{w}|}$$
$$s.t. \qquad \mathbf{w}^T \mathbf{w} = 1, \tag{2.2}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class scatter matrix and within-class scatter matrix respectively. Similarly, $\mathbf{w}^T\mathbf{w} = 1$ is only for a scaling factor and will not affect the optimization process.

A comparison between the PCA and LDA methods was conducted in references [81] [82]. Experiment results demonstrate that LDA performs better than PCA when the number of the training images per class is sufficiently large, and PCA outperforms LDA only when the number of the training data per class is limited. Although LDA illustrated a good recognition performance, a drawback cannot be ignored that the performance of LDA was often degraded due to the singularity of $\mathbf{S}_W$. To avoid this, many extensions based on the LDA method were proposed such as Nullspace LDA [83] [84], Regularized Discriminant Analysis [85], 2D-LDA [86] [87], Enhanced FLD [88], Generalized Singular Value Decomposition [89] [90] etc. All these enhanced algorithms provide better recognition rates and more robust performance than the LDA baseline.

Summary: The main drawback of the above linear dimension reduction algorithms, such as PCA, ICA and LDA, is the limited capability of linearly capturing the characteristics of the underlying structure of data that lie on a nonlinear manifold. To efficiently describe the data with nonlinear distributions, many techniques were proposed to nonlinearly learn the transformations for dimension reduction and better face representation, including Isometric Feature Mapping (ISOMAP) [91], Laplacian Eigenmap (LE) [92], Locally Linear Embedding (LLE) [93] etc. This motivated a paradigm shift from learning linear transformations to learning nonlinear transformations of face images.

- **Learning nonlinear transformations—ISOMAP, LE, LLE, etc.**

  ISOMAP: A successful attempt to recognize objects and human faces in a nonlinear manner was proposed in a manifold learning method [91] called Isometric

Feature Mapping (ISOMAP), which aims to find the underlying low-dimensional structure with nonlinear distributions of a data set. Generally, the data set is considered as a nonlinear manifold that is a topological space with Euclidean locality. In ISOMAP, the authors first constructed a neighborhood graph where the pairwise geodesic distances, i.e., the shortest distance between two sample points along the surface of a nonlinear manifold, was computed and stored. Based on this neighborhood graph, the classical Multidimensional scaling method (MDS) [94] was applied to seek an embedding that preserves the pairwise geodesic distances in the low-dimensional space. Suppose $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ is a data set in the high-dimensional space, for which the matrix of the geodesic distance graph is given as $\mathbf{D}_G(i,j) = \{d_G(\mathbf{x}_i, \mathbf{x}_j)\}$. Then the embedded data set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ can be obtained by embedding the data set $\mathbf{X}$ in a low-dimensional Euclidean space. In the low-dimensional space, the matrix of the Euclidean distance graph is given as $\mathbf{D}_E(i,j) = \{d_E(\mathbf{y}_i, \mathbf{y}_j)\}$. Thus the objective of ISOMAP is to keep the difference between the geodesic distance and the Euclidean distance of the corresponding points as small as possible after embedding, i.e.

$$\min \|\tau(\mathbf{D}_G) - \tau(\mathbf{D}_E)\|_{L_2}, \tag{2.3}$$

where τ is an operator that transforms distances to inner products, and $\|\cdot\|_{L_2}$ denotes the L_2 matrix norm.

LLE: In contrast to ISOMAP that learns an embedding by constraining the global geometric relationships, the Locally linear embedding method proposed in (LLE) [93] aims to seek the embedding through restricting only the locality. By locality, a point and its neighbors span a low-dimensional Euclidean space on the surface of a manifold. In view of this, LLE combines the global topological relationships, i.e. the connections among neighbor points, and linear locality of a manifold to learn the nonlinear embedding. In stead of computing the geodesic distances between one point and all the other points on the manifold, LLE only calculates the linear reconstruction coefficients of one point from its neighbors and stored them in a graph. The objective of LLE is to keep the coefficients unchanged after embedding,

i.e.

$$\mathbf{w}_i^* = \arg\min_{\mathbf{w}_i} \sum_i |\mathbf{x}_i - \sum_j w_{ij}\mathbf{x}_j|^2, \qquad s.t. \quad \sum_j w_{ij} = 1, \tag{2.4}$$

$$\min_{\mathbf{y}_i} \sum_i |\mathbf{y}_i - \sum_j w_{ij}^*\mathbf{y}_j|^2, \tag{2.5}$$

where $\mathbf{w}_i$ is the coefficients of the linear reconstruction of point $\mathbf{x}_i$.

LE: Similar to ISOMAP and LLE, the Laplacian eigenmap (LE) method [92] also explores the topological relationships among data to construct a graph. The core idea of LE is to find a nonlinear embedding that puts the points as close as possible if they are neighbors to each other in the original space, i.e.

$$\min \sum_{ij} \mathbf{S}_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2, \tag{2.6}$$

where $\mathbf{S}$ is the affinity graph obtained by calculating the heat kernels $\mathbf{S}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta)$ if $\mathbf{x}_j$ belongs to the k nearest neighbors of $\mathbf{x}_i$; otherwise $\mathbf{S}_{ij} = 0$. This is equivalent to solving a generalized eigen-decomposition problem

$$\min tr(\mathbf{Y}^T\mathbf{L}\mathbf{Y}), \quad s.t. \quad \mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I}, \tag{2.7}$$

where the Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ and $\mathbf{D}_{i,i} = \sum_{j=1}^{n} \mathbf{S}_{i,j}$ and $\mathbf{D}_{i,j} = 0$ for $i \neq j$. To avoid trivial solutions, LE imposes a constraint $\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I}$ on this objective function.

Summary: To exploit much more compact description of a manifold meanwhile avoid the curse of dimensionality, a nonlinear dimension reduction procedure is usually applied on a manifold. However, the main drawback of these manifold learning approaches is the poor capability of generalization due to the fact that there is no explicit mapping functions found for new input samples.

- **Using linear transformations to approximate nonlinear transformations—Graph Embedding approaches such as LPP**

LPP: Based on the proof [95] that the Laplacian of a graph can be considered as a good approximation of Laplace-Beltrami operator defined on the manifold, many graph embedding algorithms have been proposed to linearly map a manifold to a low dimensional subspace, such that the embedding is defined everywhere. Among these methods, Locality Preserving Projections (LPP) [95] is the first attempt to prove and linearize the underlying nonlinear mapping and represent it explicitly. As a good linear approximation of LE, LPP shares the similar idea and greatly extends the generalization ability of LE. Suppose that a unknown linear projection $\mathbf{w}$ (a column vector) maps the original data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ from the high-dimensional space to the projected data set $\mathbf{y} = [y_1, y_2, \cdots, y_N]$ in the low-dimensional space, i.e., $\mathbf{y} = \mathbf{w}^T\mathbf{X}$, then the generalized eigen-decomposition problem becomes

$$\min(\mathbf{w}^T\mathbf{X})\mathbf{L}(\mathbf{w}^T\mathbf{X})^T, \quad s.t. \quad (\mathbf{w}^T\mathbf{X})\mathbf{D}(\mathbf{w}^T\mathbf{X})^T = 1. \tag{2.8}$$

Through this linear transformation $\mathbf{w}$, the locality characteristic of a manifold can be optimally preserved and a new input sample can be directly projected and used for classification without re-learning.

Graph Embedding: For the purpose of dimension reduction, Yan et al. proposed a general framework known as graph embedding [96].

$$\mathbf{y}^* = \arg\min_{\mathbf{y}^T\mathbf{P}\mathbf{y}=\beta} \mathbf{y}^T\mathbf{L}\mathbf{y}, \tag{2.9}$$

where $\mathbf{L}$ is the Laplacian matrix of an intrinsic graph and $\mathbf{P}$ is a constraint matrix, which can be either a diagonal matrix or the Laplacian matrix of a penalty graph. Within this framework, some classic subspace learning approaches such as PCA, LDA, LPP, ISOMAP, LLE and LE can be considered as special cases. These algorithms have different motivations but share the common form in mathematics. Furthermore, the graph embedding framework explains the intrinsic implication of each graph so that an explicit way to design the graphs according to special purposes is provided. With different objectives, we can thus design different intrinsic graphs and penalty graphs accordingly.

- **New trends—Sparse representations**

  Recently, sparse coding has shown promising properties for face reconstruction, representation and recognition. A representative face recognition method, which utilizes the advantages of sparse representation, is the sparse representation-based classification (SRC) method [97]. By exploiting the fact that the errors caused by occlusion and corruption are often sparse in a face image, the SRC method is tolerant to these sparse errors and can achieve a more robust face recognition performance. Furthermore, the SRC method is restricted to seeking sparse coefficients such that it naturally emphasizes the bases of a class to which the test image belongs. Thus the sparse representations obtained from SRC are discriminative and suitable for classification.

It has been shown in the literature that the holistic information is crucial for humans to recognize faces. In the appearance-based holistic approaches, the global information of each face image is preserved rather than using only the pixels in certain positions or local regions. However, the holistic representation-based methods would be affected by the variations in terms of scale and misalignments, which might be better handled by the local feature-based methods.

2.3 Face Recognition Based on Image Sets

2.3.1 Motivation

- **Difficulties confronted by the conventional face recognition systems**

  As shown in Fig. 1.1 in Chapter 1, the conventional face recognition systems are usually based on image-to-image matching that treats each face image as an isolated individual. This matching mode may bring about some practical problems to the complex recognition systems that usually consist of face detection/ segmentation/ tracking, face alignment, face representation and face identification. For high quality face images such as drivers' licenses, passports and IC photos, face segmentation and recognition are rather easy procedures since the face images are acquired under a controlled condition. However, in practical scenarios there is considerable

variations in face images that are collected from a wide range of sources and settings, e.g., images contain faces in arbitrary poses, clutter surroundings with low resolutions or occlusions. Under these uncontrolled conditions, face detection and recognition will confront serious challenges. The performance of many techniques might be degraded heavily when multiple variations are present simultaneously, e.g., a pose-invariant face recognition approach may not work properly when the face images also contain variations in expression or illumination. There is no universal technique that can deal with all practical situations. Moreover, there is no evidence to prove that a complex situation can be handled by the combinations of different techniques. Thus the conventional face recognition systems under the unconstrained conditions are struggling in providing robust and reliable performance. Compared to a single-shot image, an image set is able to provide more useful information to cover more variations in the individuals' appearances. Hence, it can provide more discriminative information to model the appearances of human faces.

- **Sources of face image sets**

  Generally, the image set can be a collection of unordered images from different views or image frames harvested from a video clip, even the video clip is not continuous. As important applications of face recognition techniques, systems for surveillance, information security, and access control are widely used in our daily lives. Moreover, a large number of social media sharing websites, such as Facebook, YouTube and Flickr, have emerged with the rapidly developing internet and constantly increasing online social interactions. Everyday hundreds of millions of people post their personal photos and videos onto these social media networks. With the massive data, video sequences and multiple still images of an individual are much easier to obtain than before.

From the above analysis, we decided to focus our research on the topic of face recognition based on image set (FRBIS). Through extensive analysis and experiments, we attempt to evaluate our proposed methods in some practical environments.

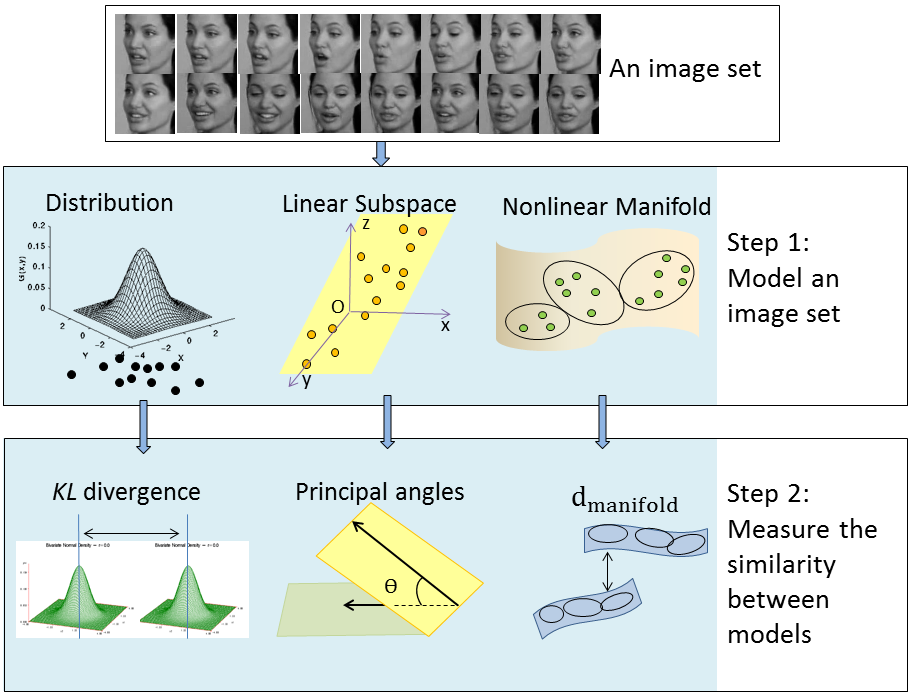

Figure 2.5: Two steps involved in the set-based classification tasks.

2.3.2 Problem Statement

Different from the conventional face recognition systems based on image-to-image matching, FRBIS attempts to recognize an individual from a set of face images. In this scenario, both the gallery and probe samples are face image sets rather than single-shot image instances and a probe set of a person needs to be classified as one of the known identities. The similarity between a gallery and a probe is obtained by calculating the set-to-set distance as shown in Fig. 1.2 in Chapter 1. As mentioned in reference [98], the task of FRBIS mainly consists of two steps: (1) the modeling of image set and (2) the similarity measuring between image sets. The modeling of image set aims to extract some features to characterize each image set and the similarity measuring between image sets aims to compute the similarity between two models of two image sets respectively, as illustrated in Fig. 2.5.

2.3.3 Related Work

There has been a growing research interest in the topic of visual recognition based on image sets where each image set contains variation information that is unavailable in a single image. As mentioned above, the FRBIS problem contains two steps, i.e., modeling an image set and measuring the similarity between such models. Once the model of an image set is designed, the corresponding similarity measure is roughly fixed. For example, the Kullback-Leibler divergence is used to measure the similarity between two statistical densities, while the principal angle is always used to measure the subspace distances. Thus, in terms of the model adopted to describe an image set, the existing FRBIS methods [99] [100] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111] [112] [113] [16] can be grouped into three categories: model-free methods, single-model based methods and multi-model based methods.

- **Model-free methods**

  As proposed in [114] [115], it is reasonable to compare two image sets by directly calculating the distances between pair-wise samples across these two sets. However, a limitation of these methods is that the recognition results may drop severely when outliers are involved in calculating the distance between two image sets.

- **Single-model based methods**

  In the single-model based methods, an image set is always formulated as a predefined parametric distribution [99] [100] [101], such as multivariate Gaussian distribution or Gaussian Mixture Models (GMM), where the parameters are estimated from the corresponding data set. Then a probability metric such as the Kullback-Leibler divergence is utilized to measure the set similarity by calculating the distance between two estimated distributions. However, the multivariate Gaussian distribution is an insufficient representation of an image set due to its insufficient ability of modeling nonlinear and low dimensional intrinsic structures [12]. In contrast, the GMM distribution is a more principled model than the Gaussian distribution and its kernel versions, but the better performance is achieved at the cost of high computational complexity from a Monte Carlo algorithm. A main

shortcoming of these parametric methods is that they may fail when there are weak statistical relationships between the training and test image sets. Furthermore, the classification performance could be affected if the practical data sets do not follow the predefined density.

To reduce the computational complexity of the face recognition systems, an image set can be modeled by using a linear subspace with low computational complexity where the missing data can be filled in, thus effectively eliminating the effects of outliers. In this scenario, the whole dataset is a collection of such linear subspaces [102] [103] [104] [105] [106] [107] [108]. One way to efficiently calculate the distance between two subspaces is utilizing the principal angles or orthonormal bases, such as Mutual Subspace Method (MSM) [105], Subspace Distance (SD) [102] and a Weighted Subspace Distance (WSD) [103]. Among these approaches, mutual subspace method (MSM) [105] is one of the first attempts to calculate the distance between two subspaces where the first (smallest) principal angle of the two subspaces was calculated as the set distance, while the discriminative information within other bases are ignored. More recently, a subspace distance (SD) method [102] was proposed to treat all the orthogonal bases of a subspace equally. However, this may result in a loss of discriminative information since the variation in each dimension is different. To prevent such a loss, a weighted subspace distance (WSD) measure [103] was proposed subsequently, which achieved a better performance than the SD method by characterizing the distribution in each dimension.

Another way to measure the similarity between subspaces is embedding a linear subspace from a Grassmann manifold to a data point in the Euclidean space through Grassmann kernels [111], based on which the distance between points in the Euclidean space can be computed as the subspace similarity. However, a major disadvantage of these subspace based methods is that their modeling abilities are insufficient to represent a nonlinear manifold. Furthermore, on a Grassmann manifold, the dimensionality of all subspaces is required to be the same.

Some novel geometric models such as affine hull and convex hull were exploited to more compactly describe an image set in the reference [98]. In these methods,

each image set was represented by a convex hull and the distance between the nearest points on the two convex hulls was calculated as the similarity between a pair of convex hulls. A support vector machine (SVM) [116] classifier was used to accomplish the set-to-set matching. However, one disadvantage of this method is its high computational complexity.

Recently, a method known as Sparse Approximated Nearest Point (SANP) [117] was proposed to measure the similarity between face image sets. In this method, each image set was modeled using an affine hull and the sparse reconstruction error was defined between two affine hulls as the distance between image sets. Although this method was able to achieve the state of the arts, the computational complexity was relatively high in solving the l1-norm objective function.

From the above analysis, a major disadvantage of the single-model based methods is the insufficient ability to represent a nonlinearly distributed image set that contains considerable variations in face images collected from a wide range of sources and settings, e.g., images contain faces in arbitrary poses, clutter surroundings with low resolutions, occlusions, etc.

- **Multi-model based methods**

  Recently, several multi-model based methods [16] [118] have been proposed to measure the similarity between two image sets by describing each image set as a nonlinear manifold. In these methods, a nonlinear manifold was usually characterized by using multiple local models. Manifold-to-Manifold distance (MMD) [16] was the first attempt to define the distance between manifolds that characterized a manifold by using several local linear patches, each of which was a local model. Owing to the unbalanced clustering method used in MMD, an enhanced version of MMD called Manifold Discriminant Analysis (MDA) [118] was proposed that modified the clustering procedures. Furthermore, a discriminative learning procedure was introduced before the manifold-to-manifold matching stage.

  The advantage of the multi-model based methods lies in the strong capability of characterizing an image set with complex distribution. However, how to efficiently compute the dissimilarity between such image sets remains a challenging problem.

2.4 Conclusion

In this Chapter, we first introduced the problem of automatic face recognition, including the advantages, problem statement, possible applications and challenges, after which a comprehensive literature review of the classic/popular methods was provided. In the rest of this thesis, we shall focus our attention on face recognition techniques based on image set (FRBIS) that was introduced in Section 2.2.

Chapter 3

Generalized Subspace Distance

3.1 Introduction

Normally, a single face image has limited capability to capture and reflect the true structure of a 3D human face, especially a low quality face image that contains a large number of variations such as occlusions, misalignments, shadows, etc. This leads to unreliable performance of the conventional face recognition techniques that recognize a person using only a single-shot image as shown in Fig. 1.1 in Chapter 1. However, a desirable property of a successful face recognition system is the capability of processing the low quality face images that are acquired and used in real-life applications. Since an image set of a person is able to provide more complete information that covers variations in an individual's appearance, one way to reduce the challenges for the current face recognition systems is to recognize a person using a set of face images instead of a single-shot image. Furthermore, a set of face images of an individual are much easier to be obtained than before due to the rapidly developing social networks. In general, a face image set of a person can be multiple shots from different viewpoints, a collection of unordered images from a personal gallery, or frames from a video clip (sometimes these frames are not in continuous sequence), as shown in Fig. 3.1.

The problem of recognizing a person based on image sets has attractive more and more research interest to achieve a practical face recognition system. As reviewed in Chapter 2, these FRBIS techniques can be grouped into three categories: model-free methods, single-model based methods and multi-model based methods. Our discussions start with the single-model based methods in this chapter, and the multi-model based

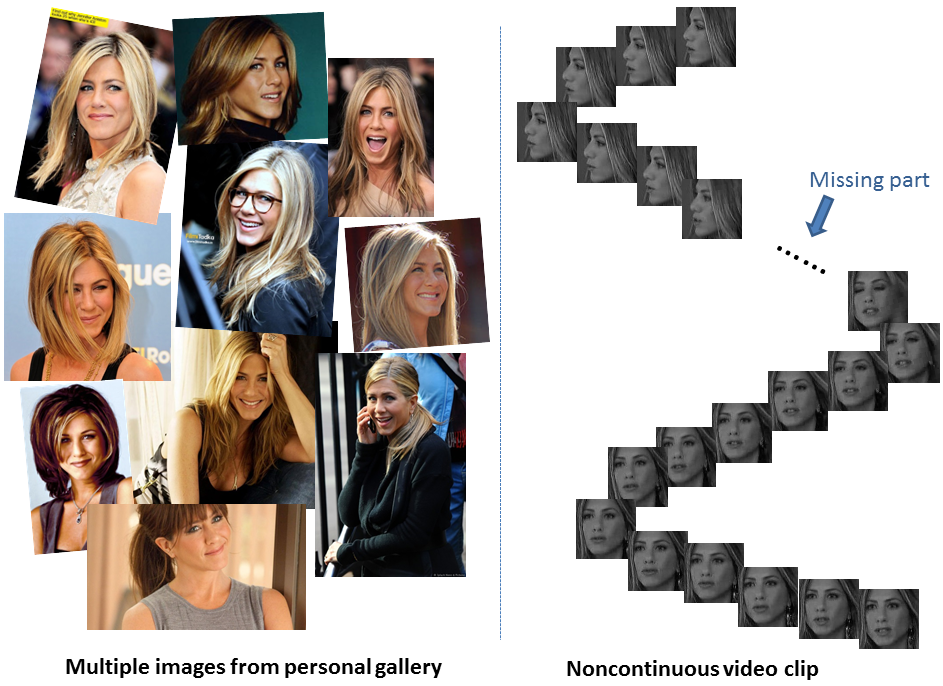

Figure 3.1: A face image set of a person might be multiple shots from different viewpoints, a collection of unordered images from a personal gallery or frames from a video clip (sometimes these frames are not in continuous sequence).

methods will be further investigated in the following chapters to compose a relatively complete study on the FRBIS problem. The model-free methods are beyond the scope of our work due to the straightforward concept and costly computational expenditures.

In the category of single-model based methods, researchers usually use a single model to describe an image set. For example, a single model can be a linear representation or a probabilistic density. One main shortcoming of the probabilistic density model is that the classification performance will be affected if the actual data sets do not follow the predefined density. Due to the limitation, the linear representation-based methods, which have been referred to as the correlation-based methods in this thesis, become more popular in image sets-based face recognition techniques. There are various linear representations proposed in the literatures, e.g., linear subspaces, affine representations and convex hulls. Among these correlation-based methods, mutual subspace method (MSM) [105] was the first attempt to model an image set by using a linear subspace where an efficient way was proposed to calculate the distance between subspaces. Following the

linear subspace model, another correlation-based method known as the subspace distance (SD) method was proposed in [102] that further explored the information of correlations among orthonormal bases for computing the distance between subspaces. As an extension of SD, a weighted subspace distance (WSD) [103] measure was proposed subsequently to improve the performance of the SD method. These correlation-based methods share a common advantage in computational efficiency, including the rapid computation of principal angles and the rapid estimation of the linear subspaces by using PCA or some incremental learning methods. In order to bring this advantage into full play, our work in this chapter focuses on how to enhance the recognition performance of the correlation-based methods.

In the rest of this chapter, we first give a brief introduction on the state of the arts, e.g., the MSM, SD, and WSD methods, where the properties and disadvantages of these methods are discussed in Section 3.2. From MSM, SD and WSD, we observe that the similarity between subspaces is always related with orthonormal bases. Thus we hope to find an explicit rule to measure the subspace distance. We observe that the weighting function of each basis is the key that leads to various subspace distances. Thus the generalized subspace distance (GSD) framework is proposed in Section 3.3 and illustrates this observation. Based on the GSD framework, in Section 3.4 we propose two distance measurements for image set matching, namely fractional order weighted subspace distance (FOWSD) and exponential weighted subspace distance (EWSD). Furthermore, the affine versions of FOWSD and EWSD are proposed in Section 3.5, which leads to a better recognition performance. In Section 3.6, extensive experiments are conducted to illustrate the effectiveness of the proposed FOWSD, EWSD and their affine versions through comparisons with the state of the arts.

3.2 Correlation-Based Methods

3.2.1 Principal Angles

To describe the relationship between two different linear subspaces, the principal angles or canonical angles, were first introduced by Hotelling in reference [112]. Suppose there are two linear subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$, and $\dim\{\mathbf{S}_1\} = \dim\{\mathbf{S}_2\} = l$. The minimal angles

$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_l \leq (\pi/2)$ between vectors of two subspaces are defined as the principal angles, for which we have

$$\begin{aligned} \cos(\theta_i) &= \max_{\mathbf{u}_i \in \mathbf{S}_1} \max_{\mathbf{v}_i \in \mathbf{S}_2} \mathbf{u}_i^T \mathbf{v}_i \\ s.t. \qquad & \mathbf{u}_i^T \mathbf{u}_i = \mathbf{v}_i^T \mathbf{v}_i = 1 \text{ and } \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0, \text{ where } j = 1, \cdots, i-1. \end{aligned} \tag{3.1}$$

The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are usually called a pair of principal vectors. The smallest angle θ_1 between two subspaces is obtained from the first pair of principal vectors $\mathbf{u}_1$ and $\mathbf{v}_1$. And the second pair of principal vectors $\mathbf{u}_2$ and $\mathbf{v}_2$, which is orthogonal to the first pair, corresponds to the second smallest angle θ_2, so on and so forth, with the next pair of principal vectors orthogonal to all the previous pairs. For ease of exposition, we use a set of orthonormal bases $\mathbf{P}_1 = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_l]$ to represent the subspace $\mathbf{S}_1$ and $\mathbf{P}_2 = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_l]$ to represent the subspace $\mathbf{S}_2$. Then an easy way to stably compute a set of principal angles between $\mathbf{S}_1$ and $\mathbf{S}_2$ can be found in reference [109], which utilized the Singular Value Decomposition (SVD)

$$\begin{aligned} (\mathbf{P}_1)^T \mathbf{P}_2 &= \mathbf{Q}_{12} \Lambda \mathbf{Q}_{21}^T \\ s.t. \qquad & \Lambda = diag(\sigma_1, \sigma_2, \cdots, \sigma_l), \end{aligned} \tag{3.2}$$

where $\mathbf{Q}_{12}$ and $\mathbf{Q}_{21}$ are orthonormal matrices, and the canonical correlations are obtained as $\cos(\theta_1) = \sigma_1, \cdots, \cos(\theta_l) = \sigma_l$. Usually, $\mathbf{P}_1$ and $\mathbf{P}_2$ are formed by eigenvectors of the data covariance matrices through a PCA algorithm.

3.2.2 Mutual Subspace Method (MSM)

Mutual Subspace Method (MSM) is the first attempt to apply the concept of principal angles to set-based face recognition techniques [105]. In this method, each image set is represented by a linear subspace and then the smallest angle between the input and the reference subspaces is used as the subspace similarity:

$$d_{MSM}(\mathbf{S}_1, \mathbf{S}_2) = \cos(\theta_1), \tag{3.3}$$

where θ_1 is the smallest principal angle obtained from (3.1).

It is noted that the MSM method uses the most similar pair of principal vectors to describe the relative positions of two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$. Due to the high tolerance

of a subspace, the performance of MSM is insensitive to the varying facial expressions and poses. Furthermore, only the first principal angle $\cos(\theta_1) = \sigma_1$ is used for computing the subspace distance, thus MSM brings an inherent advantage in the computational efficiency.

3.2.3 Subspace Distance (SD)

Although the MSM is an efficient distance metric for measuring the similarity between linear subspaces, it ignores much useful information carried by other pairs of orthonormal bases. Furthermore, the distance between subspaces becomes unreliable when the first principal angle is affected by some variations. In order to put the information carried by all the orthonormal bases of $\mathbf{S}_1$ and $\mathbf{S}_2$ into full play, a subspace distance (SD) was proposed in [102] where the subspace distance (SD) can be computed between two subspaces with different dimensions. Given an m-dimensional subspace $\mathbf{S}_1$ and an n-dimensional subspace $\mathbf{S}_2$ where m and n denote the numbers of orthonormal bases in subspace $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively, we assume $m < n$ without losing generality. Then the subspace distance (SD) is defined as

$$d_{SD}(\mathbf{S}_1, \mathbf{S}_2) = \sqrt{\max(m, n) - \sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{p}_i^T\mathbf{q}_j)^2}. \tag{3.4}$$

It is observed that all the orthonormal bases $\mathbf{p}_i$ and $\mathbf{q}_j$ are involved in the measurement of subspace distance. In [102], the subspace distance (SD) was proved to be invariant to the choice of orthonormal bases, thus the eigenvectors of data covariance matrices are usually an intuitive choice to represent subspaces.

3.2.4 Weighted Subspace Distance (WSD)

It is observed that the SD method treats all the orthogonal bases of a subspace equally. However, the variation of a data set in each dimension is different. To capture the different data distributions of an image set, a weighted subspace distance was proposed in reference [103] where each orthonormal basis is assigned with a weight

$$d_{WSD}(\mathbf{S}_1, \mathbf{S}_2) = \sqrt{1 - \sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\lambda_i\mu_j}(\mathbf{p}_i^T\mathbf{q}_j)^2}, \tag{3.5}$$

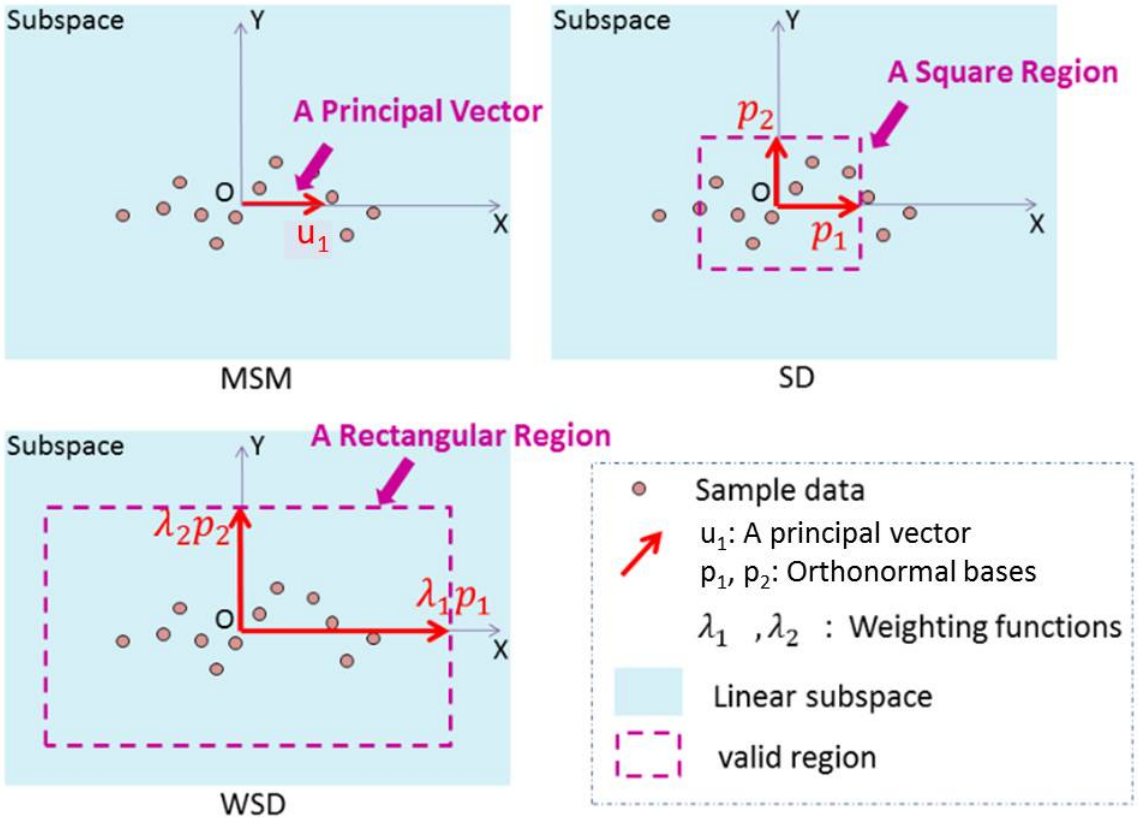

Figure 3.2: From the upper-left to lower-right: the illustrations of MSM, SD, WSD and notations. The MSM method uses a principal vector to describe an image set, the SD method uses a square region to describe an image set and the WSD method uses a rectangular region to describe an image set.

where λ_i and μ_j are the eigenvalues corresponding to the orthonormal bases $\mathbf{p}_i$ and $\mathbf{q}_j$ respectively.

3.2.5 Comparisons Between MSM, SD and WSD

As shown in Fig. 3.2, we illustrate the basic concept of the MSM, SD and WSD methods geometrically. For each method shown in Fig. 3.2, a valid vector or region is chosen to illustrate the actually used part of each subspace, i.e., the valid vector or region is the actual representation of an image set for calculating the set-to-set distance. In Fig. 3.2, the upper-left part shows that the MSM method only uses the first principal vector $\mathbf{u}_1$ as a representation of the image data set for computing the subspace distance measurement.

This is a rather weak representation that loses most information of an image set. In contrast, the SD method exploits all the orthonormal bases, which constructs a valid square region for measuring the similarity between image sets, as shown in the upper-right part of Fig. 3.2. However, We can observe that the square region used in the SD method does not successfully capture the actual distribution of an image set due to the fact that SD treats all the orthonormal bases equally. To characterize the distribution of an image set more reasonably, a rectangular region is used in the WSD method, which assigns eigenvalues to the corresponding bases. It is observed that the weighting functions of WSD are eigenvalues that are fixed values and cannot be changed even if it is not optimal for classification.

3.3 Proposed Generalized Subspace Distance (GSD) Framework

From the above analysis, all the distance measures MSM, SD and WSD are designed for measuring the dissimilarity between two subspaces by using the orthonormal bases. Next, we summarize the common properties of these measures and propose a generalized subspace distance (GSD) framework, within which the SD and WSD distance measures can be considered as special cases.

As defined in Section 3.2, we use $\mathbf{P}_1 = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m]$ to represent the subspace $\mathbf{S}_1$ and $\mathbf{P}_2 = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n]$ to represent the subspace $\mathbf{S}_2$ respectively and we let $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ and $\{\mu_1, \mu_2, \cdots, \mu_n\}$ be the corresponding eigenvalues of $\mathbf{P}_1$ and $\mathbf{P}_2$.

Taking into account the different importance of the bases, all the orthogonal bases are utilized by the proposed generalized subspace distance (GSD) to calculate the subspace distance

$$d_{GSD}(\mathbf{S}_1, \mathbf{S}_2) = \sqrt{1 - \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{f(\lambda_i)}{\sum_{i=1}^{m} f(\lambda_i)}\right]^{1/2}\left[\frac{f(\mu_j)}{\sum_{j=1}^{n} f(\mu_j)}\right]^{1/2}(\mathbf{p}_i^T\mathbf{q}_j)^2}, \tag{3.6}$$

where each of the bases is assigned with a weighting function $f(\lambda_i)$ and $f(\mu_j)$, $1 \le i \le m$ and $1 \le j \le n$, such that an unique weight is produced for each basis. The weighting function can be flexibly designed for different purposes.

It is obvious that the key procedure of the GSD is how to design a weighting function that makes the GSD a valid distance metric and better for classification. With different weighting functions f, the existing subspace distance measures such as SD and WSD can be readily derived from our proposed GSD framework as following:

(1) when $f = 1 \; \forall \; i, j$, the proposed GSD degenerates into the conventional SD, which assigns *equal weights* to all bases;

(2) when $f(\lambda_i) = \lambda_i$ and $f(\mu_j) = \mu_j$, $\forall \; i, j$, the proposed GSD degenerates into the WSD;

As a valid distance metric, the proposed GSD also needs to satisfy several properties such as nonnegativity, symmetry, triangle inequality, etc. Most of these properties are straightforward to justify, thus we only prove the validity and triangle inequality as below.

- **Validity**:

  Proof: Let $\{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m\}$ and $\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n\}$ be the orthonormal bases of subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively, then we have

$$\sum_{j=1}^{n} (\mathbf{p}_i^T \mathbf{q}_j)^2 \leq \|\mathbf{p}_i\|^2 = 1, \tag{3.7}$$

$$\sum_{i=1}^{m} (\mathbf{p}_i^T \mathbf{q}_j)^2 \leq \|\mathbf{q}_j\|^2 = 1. \tag{3.8}$$

  For ease of exposition, we let $\widetilde{f}(\lambda_i) = \frac{f(\lambda_i)}{\sum_{i=1}^{m} f(\lambda_i)}$ and $\widetilde{f}(\mu_j) = \frac{f(\mu_j)}{\sum_{j=1}^{n} f(\mu_j)}$, then we have

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} \sqrt{\widetilde{f}(\lambda_i)\widetilde{f}(\mu_j)} (\mathbf{p}_i^T \mathbf{q}_j)^2 &\leq \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\widetilde{f}(\lambda_i) + \widetilde{f}(\mu_j)}{2} (\mathbf{p}_i^T \mathbf{q}_j)^2 \\
&= \frac{1}{2}\sum_{i=1}^{m} \widetilde{f}(\lambda_i) \sum_{j=1}^{n} (\mathbf{p}_i^T \mathbf{q}_j)^2 + \frac{1}{2}\sum_{j=1}^{n} \widetilde{f}(\mu_j) \sum_{i=1}^{m} (\mathbf{p}_i^T \mathbf{q}_j)^2 \\
&\leq \frac{1}{2}\sum_{i=1}^{m} \widetilde{f}(\lambda_i) + \frac{1}{2}\sum_{j=1}^{n} \widetilde{f}(\mu_j) = 1.
\end{aligned} \tag{3.9}$$

  Thus, d_{GSD} is a valid distance metric provided that the radicand is nonnegative.

- **Triangle inequality**: Next, we prove the triangle inequality of d_{GSD} in a similar way to the references [102] [103].

  Proof: Let $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ be m, n and k-dimensional subspaces of $\Re^D$. Let $[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m]$ be the orthonormal bases of $\mathbf{P}$, $[\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n]$ be the orthonormal bases of $\mathbf{Q}$ and $[\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_k]$ be the orthonormal bases of $\mathbf{R}$. Then we construct $D \times D$ matrix $\mathbf{M}_P$:

$$\mathbf{M}_P = \sum_{i=1}^{m} \mathbf{M}_P(i) = \sum_{i=1}^{m} \sqrt{\widetilde{f}(\lambda_i)}(\mathbf{p}_i \mathbf{p}_i^T). \tag{3.10}$$

  Since

$$\begin{aligned} \|\mathbf{M}_P(i)\|_F^2 &= tr\{\mathbf{M}_P^T(i)\mathbf{M}_P(i)\} \\ &= \widetilde{f}(\lambda_i) tr\{\mathbf{p}_i \mathbf{p}_i^T \mathbf{p}_i \mathbf{p}_i^T\} \\ &= \widetilde{f}(\lambda_i), \end{aligned} \tag{3.11}$$

  and for $i \neq j$:

$$tr\{\mathbf{M}_P^T(i)\mathbf{M}_P(j)\} = \sqrt{\widetilde{f}(\lambda_i)\widetilde{f}(\lambda_j)} tr\{\mathbf{p}_i \mathbf{p}_i^T \mathbf{p}_j \mathbf{p}_j^T\} = 0, \tag{3.12}$$

  thus we have

$$\|\mathbf{M}_P\|_F^2 = \sum_{i=1}^{m} \|\mathbf{M}_P(i)\|_F^2 = \sum_{i=1}^{m} \widetilde{f}(\lambda_i) = 1. \tag{3.13}$$

  Similarly, we have $\|\mathbf{M}_Q\|_F^2 = \|\mathbf{M}_R\|_F^2 = 1$. Then the Frobenius norm of $(\mathbf{M}_P - \mathbf{M}_Q)$ can be derived as

$$\begin{aligned} \|\mathbf{M}_P - \mathbf{M}_Q\|_F &= \sqrt{\|\mathbf{M}_P\|_F^2 + \|\mathbf{M}_Q\|_F^2 - 2 \cdot tr(\mathbf{M}_P^T \mathbf{M}_Q)} \\ &= \sqrt{2 - 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\widetilde{f}(\lambda_i)\right]^{1/2} \left[\widetilde{f}(\mu_j)\right]^{1/2} (p_i^T q_j)^2} \\ &= \sqrt{2} \cdot d_{GSD}(\mathbf{P}, \mathbf{Q}). \end{aligned} \tag{3.14}$$

  Thus, it is obvious that

$$\|\mathbf{M}_P - \mathbf{M}_Q\|_F \leq \|\mathbf{M}_P - \mathbf{M}_R\|_F + \|\mathbf{M}_R - \mathbf{M}_Q\|_F \tag{3.15}$$

$$\Longrightarrow d_{GSD}(\mathbf{P}, \mathbf{Q}) \leq d_{GSD}(\mathbf{P}, \mathbf{R}) + d_{GSD}(\mathbf{R}, \mathbf{Q}). \tag{3.16}$$

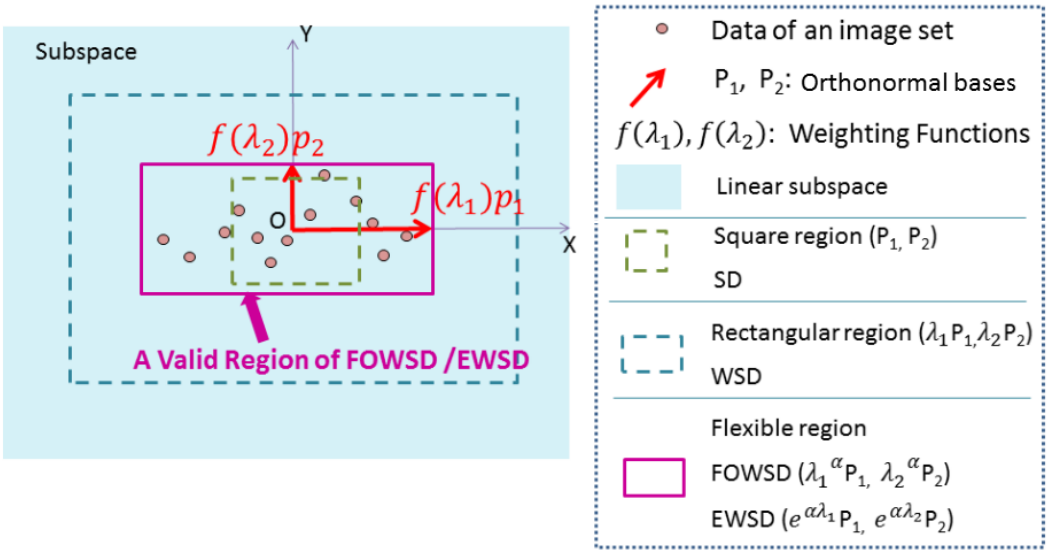

Figure 3.3: The rose-colored region is used in the FOWSD and EWSD methods to approximate an image set for calculating the set-to-set distances. Compared with the vector and regions used in SD and WSD, the rose-colored region is more suitable and flexible.

3.4 Proposed FOWSD and EWSD

It is observed that the weighting function f in the GSD framework is what distinguishes one distance measure from another and is probably the key to further improving the classification performance. Based on this GSD framework, we design some simple but effective distance metrics to further improve the performance of the conventional SD and WSD. By introducing a parameter α that is obtained from a learning stage, a fractional order weighted subspace distance (FOWSD) and an exponential weighted subspace distance (EWSD) are proposed to discriminatively measure the distance between image sets. Compared with the SD and the WSD methods where the valid regions do not fit an image set appropriately, the parameter α enables FOWSD and EWSD to use a more flexible and compact valid region to approximate an image set, as illustrated in Fig. 3.3.

3.4.1 Weighting Functions

Based on the geometric concept illustrated in Fig. 3.3, we define the weighting functions of FOWSD as

$$f(\lambda_i) = \lambda_i^{\alpha}, \tag{3.17}$$

$$f(\mu_j) = \mu_j^{\alpha}. \tag{3.18}$$

As will be shown later in the experiments, the optimal α always reside between 0 and 2, with a sharp deterioration below 0 and a more moderate one above 2. With this observation, we set a finite range for α to be searched in, i.e., $\alpha \in [0, 2]$, which makes the proposed distance take fractional ordered weights. Similarly, we define the weighting functions of EWSD as

$$f(\lambda_i) = e^{\alpha\lambda_i}, \tag{3.19}$$

$$f(\mu_j) = e^{\alpha\mu_j}, \tag{3.20}$$

where $1 \leq i \leq m$ and $1 \leq j \leq n$, and α is a parameter that is obtained in a learning stage from the training data. Similar to the WSD method, the original terms λ_i^{α}, μ_j^{α}, $e^{\alpha\lambda_i}$ and $e^{\alpha\mu_j}$ are normalized for constructing an image brightness-invariant distance in (3.6). This distance makes use of the eigenvalues to reflect the sample variance along different dimensions. The spread of the classes within their respective subspaces can be manipulated by adjusting α. Thus, a more suitable and flexible region is obtained in the FOWSD and EWSD methods to measure the image set-to-set dissimilarity.

3.4.2 Finding the Optimal Parameter

To determine an appropriate value for the parameter α, we have designed a discriminative learning procedure through which the within-class distances among image sets from the same class can be reduced. In conventional discriminant analysis algorithms, the within-class compactness is usually characterized by the average of all within-class distances. However, minimizing this average distance cannot guarantee the compactness for every class due to the heterogeneities of different classes. In order to make the within-class distance of every class as small as possible, we have designed a discriminative learning

method to search the optimal parameter α by minimizing the maximum within-class distance for FOWSD and EWSD, as shown in Fig. 3.4.

We assume that there are C training image sets from C classes respectively. We randomly divide all the available training images in each class into two subsets, which are characterized by subspaces s^1 and s^2 respectively by applying PCA. Hence, s_i^1 denotes one subset from the i-th class and s_i^2 denotes the other subset from the i-th class, where $i \in \{1, 2, \cdots, C\}$. Then the within-class distance between these two subsets is defined as

$$D_w(i, \alpha) = d_{FOWSD}(s_i^1, s_i^2), \quad i \in \{1, \cdots, C\}, \tag{3.21}$$

which indicates the compactness of the class i. Instead of minimizing the average over all within-class distances, we attempt to minimize the maximum within-class distance over all classes. Then the optimal α^* is given as

$$\alpha^* = \arg\min_{\alpha} \max_{i} D_w(i, \alpha), \quad i \in \{1, \cdots, C\}. \tag{3.22}$$

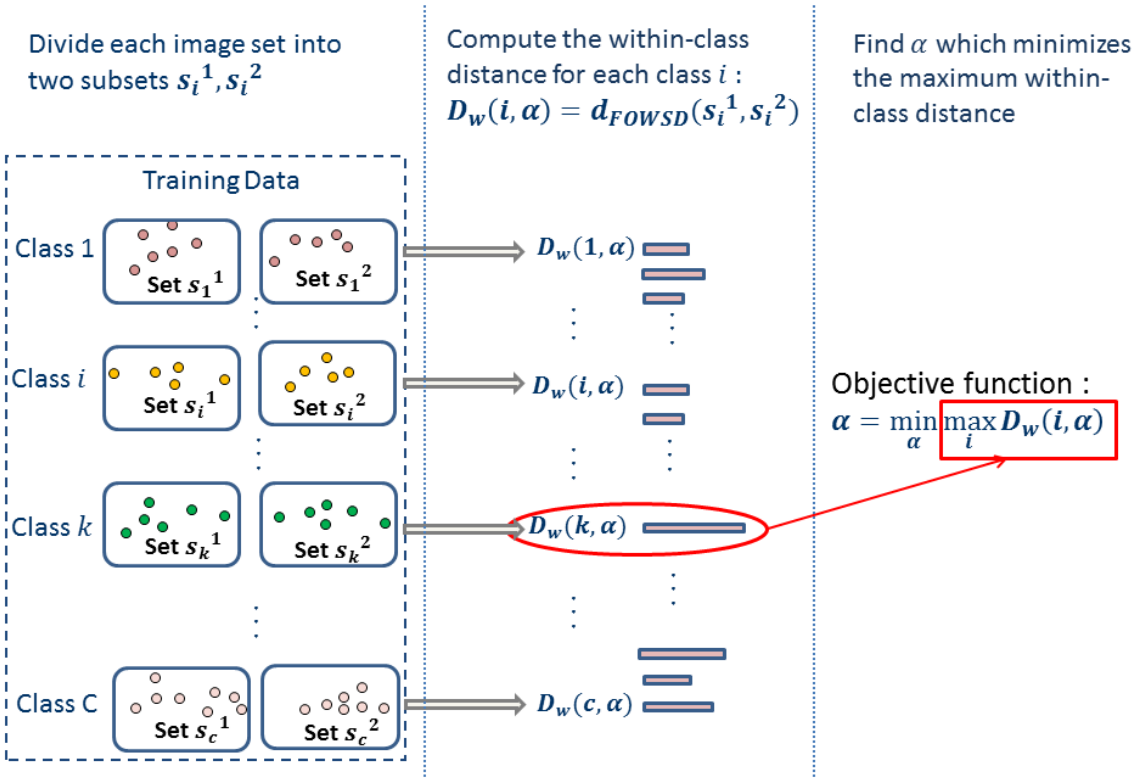

Figure 3.4: The three steps for discriminatively searching the optimal parameter α.

With the optimal α^*, the compactness of all classes can be attained.

However, through some preliminary experiments, we observed that the solution of the objective function defined in (3.22) does not always lead to the best performance of classification tasks due to disturbances, such as outliers and multiple instances with the same maximum within-class distance. To overcome this problem and improve the robustness of our method, we define the maximum within-class distance $D_w(k_1, \alpha)$ and the second maximum within-class distance $D_w(k_2, \alpha)$ respectively

$$D_w(k_1, \alpha) = \max_i D_w(i, \alpha), \quad i \in \{1, \cdots, C\}, \tag{3.23}$$

$$D_w(k_2, \alpha) = \max_i D_w(i, \alpha), \quad i \in \{1, \cdots, C\} \text{ and } i \neq k_1. \tag{3.24}$$

Hence, our objective function is to minimize the *sum* of the *two maximum* within-class distances

$$\alpha^* = \min_\alpha \left[D_w(k_1, \alpha) + D_w(k_2, \alpha) \right]. \tag{3.25}$$

Since the maximization in (3.23) and (3.24) is over discrete variable i, the minimization in (3.25) is discontinuous and thus cannot be solved analytically. In our preliminary experiments, we have observed that the optimal α always resides between 0 and 2, with a sharp deterioration below zero and a more moderate one above 2. In view of this, we set a finite range for α to be searched in, i.e., $\alpha \in [0, 2]$, which makes our proposed distance take fractional ordered weights. In this given range, an optimal α can be obtained by enumerating the sum of the two maximum within-class distance defined in (3.25) with a small but finite step size. If the search were to result in an optimal α around the boundary, we would further extend the search range in anticipation of a better α beyond. The effectiveness of this discriminative learning procedure will be demonstrated later in Section 3.6.

3.5 Proposed Affine-FOWSD and Affine-EWSD

We have observed from experiments that when we perform PCA on each image set to obtain its orthonormal bases, the SVD algorithm is actually performed after each image set is centered. This means that we lose one dimension information (the mean vector) of each image set. Thus we have made a little modification to the FOWSD and

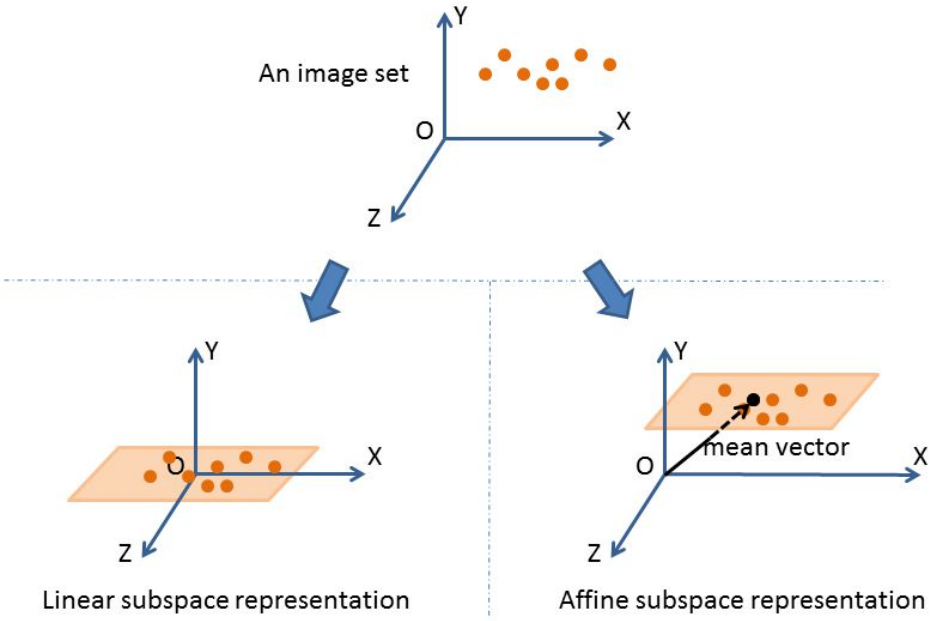

Figure 3.5: The upper subfigure shows an image set. The bottom-left subfigure is the linear subspace representation of this image set. The orthonormal bases of the linear subspace are obtained from PCA where SVD is usually performed after data centered. The bottom-right subfigure is the affine subspace representation of this image set, which contains orthonormal bases and a mean vector.

EWSD by adding the mean vector of each image set, and the resulted subspace distances are denoted as Affine-FOWSD and Affine-EWSD since they can be considered as the distances between affine subspaces. We illustrate the idea in Fig. 3.5.

The upper part of Fig. 3.5 shows an image set. In the bottom-left subfigure, the $X - Z$ plane is the linear subspace representation of this image set. Generally, the orthonormal bases of a linear subspace can be obtained using PCA. As a common practice in PCA, SVD is usually performed after the image set is centered. Thus the one-dimension information (the mean vector) is discarded by the linear subspace representation. In the bottom-right subfigure, the flat surface parallelled to the $X-Z$ plane is the affine subspace representation of this image set, which contains orthonormal bases and a mean vector. Hence, due to the affine representations, the proposed Affine-FOWSD and Affine-EWSD are more discriminative for classification.

Suppose there are two image sets $\mathbf{S}_1$ and $\mathbf{S}_2$, and $\mathbf{m}_1$ and $\mathbf{m}_2$ denote the mean vectors of $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively. Following the idea illustrated in Fig. 3.5, we define the

Affine-FOWSD as

$$d_{FOWSD}^{aff} = \beta\|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + (1-\beta)\cdot d_{FOWSD}(\mathbf{S}_1, \mathbf{S}_2). \tag{3.26}$$

Similarly, the Affine-EWSD is defined as

$$d_{EWSD}^{aff} = \beta\|\mathbf{m}_1 - \mathbf{m}_2\|_2^2 + (1-\beta)\cdot d_{EWSD}(\mathbf{S}_1, \mathbf{S}_2). \tag{3.27}$$

The principal vectors of set $\mathbf{S}_1$ can be considered as a coordinate system with respect to its mean vector $\mathbf{m}_1$, the same applies for set $\mathbf{S}_2$. Thus, the term $\|\mathbf{m}_1 - \mathbf{m}_2\|_2^2$ in (3.26) and (3.27) quantifies how far away the origins of the two coordinate systems are from each other. And the term d_{FOWSD} or d_{EWSD} indicates that when the centers of the two coordinate systems are both shifted to the same origin, how relevant the two linear subspaces are. Then β is a parameter for balancing the importance between the two terms.

3.6 Experiments

In this section, we conduct experiments towards the following objectives: (1) Illustrating the effectiveness of the proposed discriminative learning procedure that is designed to search the optimal parameter α; (2) Evaluating the performance of the proposed FOWSD and EWSD and comparing them with the state of the arts to demonstrate the effectiveness; (3) Evaluating the performance of the proposed Affine-FOWSD and Affine-EWSD and comparing them with FOWSD and EWSD to show the improvements brought by the affine representations.

3.6.1 Experimental Settings

In the following experiments, two image classification tasks, i.e., object recognition and face recognition, are conducted on the CMU MoBo database [119], the ALOI database [120], the YaleB database [121], and the ETH80 database [122], respectively.

ALOI database: The Amsterdam Library of Object Images (ALOI) database [120] is a collection of $110,250$ images captured from 1000 different objects, including toys, Rubik's cube, cups, clocks, etc. From the first image of each object, a fixed incremental

Figure 3.6: The object images inside an image set from the ALOI database.

value of variations in viewing angle, illumination angle and color, is uniformly added to the next image. In our experiments, we use the gray version of this database and randomly select 200 objects from the 1000 objects to conduct our evaluations. In the selected subset of ALOI, each image set of an object contains 72 images in different viewing angles $(0°, 5°, \cdots, 355°)$. In the experiments, we resize each image to the size of 32 by 24 pixels. An image set from ALOI database is illustrated in Fig. 3.6.

ETH80 database: The ETH80 database, which contains 3280 images, was created by Bastian Leibe and Bernt Schiele in reference [122] for object categorization. There are 80 objects from 8 categories. For each category, high-resolution color images of 10 objects are carefully collected that contain large within-class variations of this category. For each object, 41 images were captured from evenly varying viewing angles. For each image, a high-quality segmentation mask was also provided to facilitate the comparisons between appearance and contour based methods. In our experiments, we resize each image to the size of 64 by 64 pixels. Some exemplar images of the objects from each category in the ETH80 database are illustrated in Fig. 3.7, which is drawn based on the original paper [122].

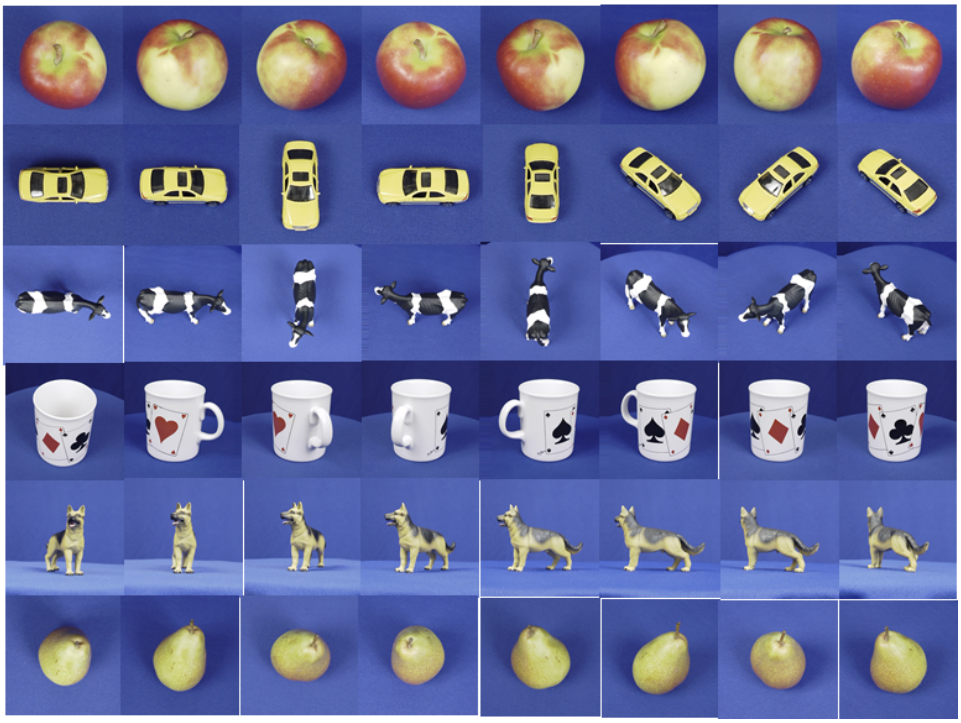

Figure 3.7: Some example object images from the ETH80 database.

The YaleB database: The Yale Face Database B (YaleB database) [121] is a collection of face images from 38 different individuals under 579 viewing conditions, including 9 different poses combined with 64 different illumination conditions. Following the experiment configurations adopted in the WSD method, we only select all the frontal face images of each person from the YaleB database, where 64 face images are selected for each of the 38 individuals. Then we resize each image to the size of 32 by 32 pixels. Some exemplar face images from the YaleB database are shown in Fig. 3.8.

CMU MoBo database: The CMU Motion of body (MoBo) database [119] was created for biometric identification of humans based on the characteristics of subjects, which contains 96 video sequences from 24 different persons. For each person, there are 4 video clips corresponding to four walk patterns, i.e., slow walk, fast walk, incline walk and walk with a ball. This results in a large number of variations in pose and expression in the face images. By applying the Viola-Jones algorithm [47], the gray scale face images can be detected and cropped to the size of 40 by 40 pixels as described in [117]. Some exemplar face images from the CMU MoBo database are shown in Fig. 3.9. Then each

Figure 3.8: Some exemplar face images from the YaleB database.

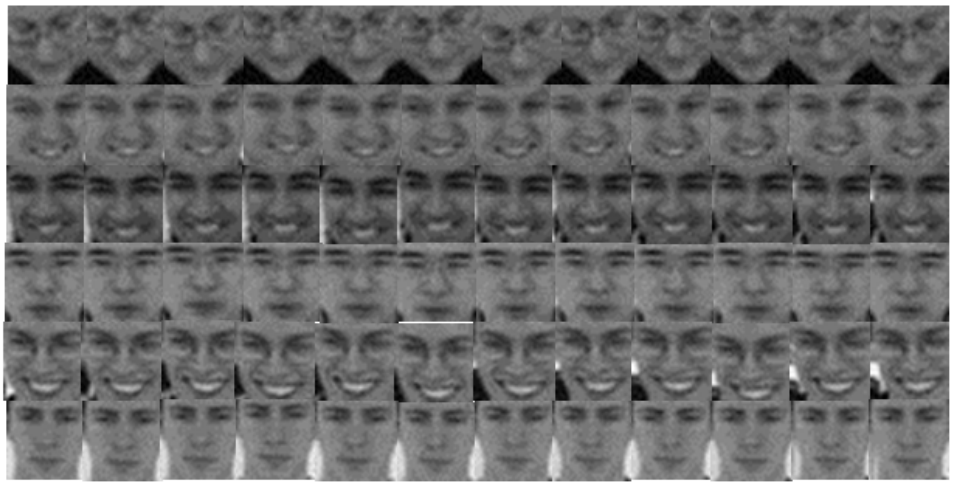

Figure 3.9: Some exemplar face images from the CMU MoBo database.

histogram equalized face image is divided into patches of size 8 by 8 pixels and a uniform Local Binary Pattern feature can be extracted from each patch as the local descriptor.

Methods for comparisons: For the state of the arts of the correlation-based methods, i.e., the MSM, SD, and WSD methods, will be evaluated and compared with our proposed methods in the following experiments. To ensure fair comparisons, the principal component analysis (PCA) algorithm is performed to each image set to obtain a collection of orthonormal bases. Based on these bases, MSM, SD, WSD, and the proposed

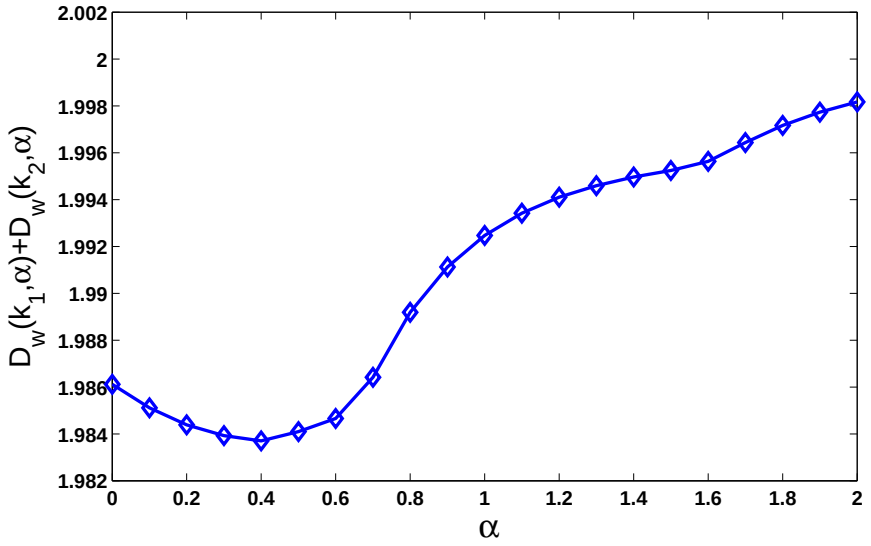

Figure 3.10: The relationship between $(D_w(k_1, \alpha) + D_w(k_2, \alpha))$ and α.

FOWSD and EWSD are performed to calculate the set-to-set distance. Then the gallery set with the highest matching core is considered as the best match with the probe set.

3.6.2 Finding the Optimal Parameter

In this experiment, we select a small subset of the ALOI database on which the proposed FOWSD is conducted as an exemplar to demonstrate the effectiveness of the proposed discriminative learning procedure that is designed to search the optimal parameter α. We randomly select 200 objects from the ALOI database and each object consists of four image sets, whose segregation arises naturally through data collection. Two image sets of each object are used for learning the parameter α (training) and the remaining two sets are used for evaluating the recognition performance (testing). In each image set, 8 images are selected and each of them is aligned and scaled to 20×20 pixels.

Applying the minimization procedures introduced earlier in Section 3.4.2 by using a step-size of 0.05, we first present the relationship between the sum of the two maximum within-class distances $(D_w(k_1, \alpha) + D_w(k_2, \alpha))$ and the parameter α in Fig. 3.10. It is apparent that $\alpha = 0.4$ corresponds to the minimum value of $(D_w(k_1, \alpha) + D_w(k_2, \alpha))$. Thus the optimal value of α learned from the minimization procedure is 0.4.

To test the optimality of the α learned from the previous step, then we use the remaining two image sets of each object to perform the object recognition task. The relationship between the recognition performance and the parameter α is shown in Fig. 3.11. It is obvious that the highest recognition rate is also achieved at $\alpha = 0.4$. This result verifies that the optimal α obtained through the proposed discriminative learning procedure enables the FOWSD method to achieve the best classification performance.

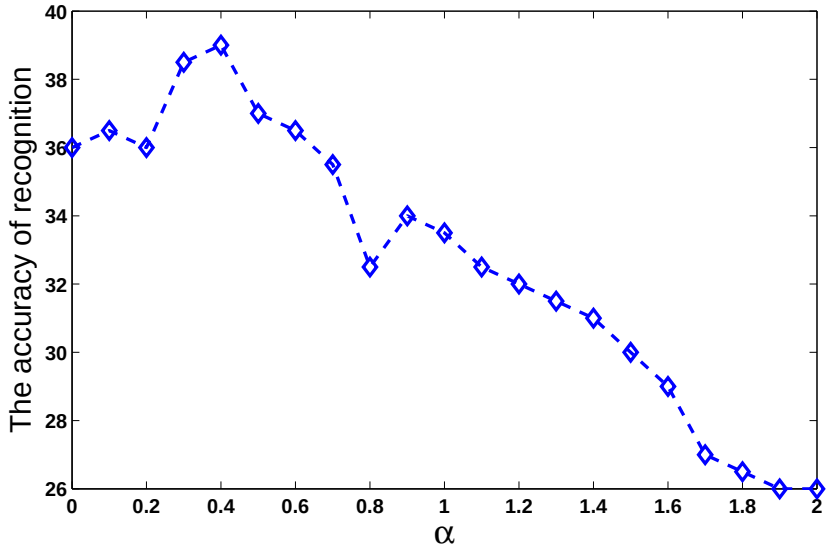

Figure 3.11: The relationship between the recognition rate and α.

3.6.3 Experimental Results of FOWSD and EWSD

We evaluate the performance of the proposed FOWSD and EWSD methods and compare them with the MSM, SD and WSD methods on four databases, i.e., the CMU MoBo database [119], the ALOI database [120], the YaleB database [121] and the ETH80 database [122], for face recognition and object recognition tasks respectively.

Face Recognition on the YaleB Database: There are 38 face image sets from 38 distinguishing in the YaleB database where each image set contains all the 64 near frontal face images. We divide each image set into 4 non-overlapped subsets. Each time we randomly select one subset per person as the gallery set and another subset per person as the probe set. Thus there are altogether 6 combinations in selecting the gallery and

probe sets for each person. The recognition results averaged over all these combinations are shown in Table 3.1.

Table 3.1: Face recognition results of the five comparable methods on the YaleB database.

Methods	Average Recognition Rate
MSM	$94.30\% \pm 1.06\%$
SD	$98.25\% \pm 0.10\%$
WSD	$87.72\% \pm 6.31\%$
Proposed FOWSD	$\mathbf{99.12}\% \pm 0.00\%$
Proposed EWSD	$98.68\% \pm 0.00\%$

It is noticed that the largest principal component in YaleB database is caused by the illumination variations in all subjects and it has nothing to do with the identity of the subjects. Thus we remove the first principal component from each image set in this experiment and obtain the experimental results in Table 3.1. From this table, it is observed that the proposed FOWSD method achieves the best recognition performance (99.12%), which outperforms the proposed EWSD (98.68%) and all other three existing methods. The improvements achieved by FOWSD over EWSD, WSD, SD, and MSM are 0.44, 11.40, 0.87 and 4.82 percentage points, respectively.

Face Recognition on the CMU MoBo Database: There are 4 face image sets for each of the 24 different persons in the CMU MoBo database. For each person, we randomly select two image sets, one is used as the gallery set and the other is used as the probe set. Similar to the configurations on the YaleB database, there are altogether 6 combinations of the gallery and probe sets. Then we illustrate the recognition rates averaged over all these 6 combinations in Table 3.2.

We can observe from Table 3.2 that the proposed FOWSD achieves the best recognition performance (93.75%) that outperforms all the other four reference methods. The improvements achieved by FOWSD over EWSD, WSD, SD, and MSM are 2.08, 2.78, 2.08 and 2.78 percentage points, respectively.

Object Recognition on the ETH80 Database: The ETH80 database contains object images from 8 categories and in each category, there are 10 image sets from 10 objects. Each time we randomly select one image set per category as the gallery set and

Table 3.2: Face recognition results of the five comparable methods on the CMU MoBo database.

Methods	Average Recognition Rate
MSM	$90.97\% \pm 0.65\%$
SD	$91.67\% \pm 0.14\%$
WSD	$90.97\% \pm 0.58\%$
Proposed FOWSD	$\mathbf{93.75}\% \pm 0.26\%$
Proposed EWSD	$91.67\% \pm 0.14\%$

Table 3.3: Object recognition results of the five comparable methods on the ETH80 database.

Methods	Average Recognition Rate
MSM	$72.08\% \pm 2.44\%$
SD	$80.00\% \pm 1.47\%$
WSD	$74.58\% \pm 2.53\%$
Proposed FOWSD	$79.17\% \pm 1.73\%$
Proposed EWSD	$\mathbf{84.58}\% \pm 1.26\%$

select another image set per category as the probe set. We repeat this process 30 times and summarize the averaged recognition results in Table 3.3.

From Table 3.3, it is observed that the proposed EWSD method achieves the best recognition rate (84.58%), which outperforms the proposed FOWSD (79.17%) and all other three methods. The improvements achieved by FOWSD over EWSD, WSD, SD, and MSM are 5.41, 10.00, 4.58 and 12.50 percentage points, respectively. From Table 3.3, we can see that the proposed FOWSD and EWSD methods are also effective in general object classification tasks.

Object Recognition on the ALOI Database: In this experiment, we randomly select 200 different objects from the ALOI database. For each object, the 72 object images are divided into 4 subsets following the same way as mentioned in reference [103]. Hence, each object contains four subsets that are characterized by the subspaces learned by PCA. Again, each time we randomly select two subsets per object, one of them is used as the gallery set and the other is used as the probe set. Similarly, there are altogether 6 combinations and the recognition results averaged over all these combinations are shown in Table 3.4.

Table 3.4: Object recognition results of the five comparable methods on the ALOI database.

Methods	Average Recognition Rate
MSM	$51.25\% \pm 2.61\%$
SD	$48.58\% \pm 1.13\%$
WSD	$49.12\% \pm 1.5\%$
Proposed FOWSD	$\mathbf{55.79}\% \pm 1.54\%$
Proposed EWSD	$52.75\% \pm 1.93\%$

Similarly, we can observe from Table 3.4 that the proposed FOWSD achieves the best recognition rate (55.79%), which outperforms all the other four methods with performance gains of 3.04, 6.67, 7.21 and 4.54 percentage points, respectively. On the other hand, it is observed that the performance of all the five methods is around 50% on the ALOI database. This is reasonable as each image set is divided into four non-overlapping subsets, such that each subset contains totally different poses from the other three subsets.

Discussions on the Experimental Results:

- The key reason that the proposed FOWSD and EWSD consistently outperform the other methods is an optimal parameter α obtained through a discriminative learning procedure. As described in the previous sections, the parameter α is obtained by minimizing the within-class distances, making the sets within a class closer and the margin between classes wider by contrast. Thus adding a new parameter α into the subspace distance measure better exploits the discriminative information, which in return improves the classification performance.

- It is observed from the experimental results that the recognition performance of the SD method is better than that of the WSD method at most of the time. Although the WSD method provides us an efficacy way to exploit more data information based on the SD method, the performance of the WSD is sometimes not as good as expected. A possible reason is that the eigenvalues are not the optimal weights of the principal vectors for the set-based image classification task. This motivates us to investigate what type of coefficients are suitable to be assigned to each principal vector as a weight and we propose a discriminative learning procedure for searching the optimal weight for each principal vector.

- Although a learning procedure is introduced in the proposed FOWSD and EWSD methods to obtain a suitable parameter α, this learning stage can be carried out off-line, without increasing the testing time required for classifying each probe image set.

3.6.4 Experimental Results of Affine-FOWSD and Affine-EWSD Methods

In this experiment, we evaluate the performance of the proposed Affine-FOWSD and Affine-EWSD methods. To illustrate the performance improvements brought by the affine subspaces, we compare the proposed Affine-FOWSD and Affine-EWSD methods with the existing MSM, SD and WSD methods on two databases, i.e., the CMU MoBo database [119] and the ALOI database [120].

Face Recognition on the CMU MoBo Database: The experimental configurations here are similar to the previous experiments where the FOWSD and EWSD methods were tested on the CMU MoBo database. Then the recognition results from all 6 combinations are averaged and summarized in Table 3.5.

Table 3.5: Face recognition results of the five comparative methods on the CMU MoBo database.

Methods	Average Recognition Rate
MSM	$90.97\% \pm 0.65\%$
SD	$91.67\% \pm 0.14\%$
WSD	$90.97\% \pm 0.58\%$
Proposed FOWSD	$\mathbf{93.75}\% \pm 0.26\%$
Proposed EWSD	$91.67\% \pm 0.14\%$
Proposed Affine-FOWSD	$\mathbf{93.75}\% \pm 0.24\%$
Proposed Affine-EWSD	$93.06\% \pm 0.32\%$

From Table 3.5, it is observed that both the proposed FOWSD and the Affine-FOWSD achieve the best recognition rate of 93.75%. The recognition performance of the Affine-EWSD reaches 93.06% that outperforms its corresponding original version 91.67% by 1.39 percentage points.

Object Recognition on the ALOI Database: The experimental configurations here are similar to the previous experiments where the FOWSD and EWSD methods were

tested on the ALOI database. Then the averaged recognition results are summarized in Table 3.6.

Table 3.6: Object recognition results of the five comparative methods on the ALOI database.

Methods	Average Recognition Rate
MSM	$51.25\% \pm 2.61\%$
SD	$48.58\% \pm 1.13\%$
WSD	$49.12\% \pm 1.5\%$
Proposed FOWSD	$55.79\% \pm 1.54\%$
Proposed EWSD	$52.75\% \pm 1.93\%$
Proposed Affine-FOWSD	$\mathbf{59.75}\% \pm 0.71\%$
Proposed Affine-EWSD	$\mathbf{59.75}\% \pm 0.60\%$

In Table 3.6, both the proposed Affine-FOWSD and Affine-EWSD methods achieve the best recognition performance of 59.75%, outperforming the original FOWSD and EWSD by 3.96 and 7.00 percentage points, respectively.

3.7 Conclusion

In this chapter, we investigated the single-model based methods for the face recognition problems based on image sets. We first introduced the motivations and the organization of this chapter in Section 3.1. For comparison purposes, the MSM, SD and WSD methods were discussed in Section 3.2 to illustrate the advantages and disadvantages of the existing correlation-based methods. We noticed that most of the existing methods failed to take into account the different importance of the individual basis of each subspace. To consolidate this family of correlation-based measures, we proposed a generalized subspace distance (GSD) framework in Section 3.3 and showed that most existing subspace similarity measures can be considered as its special cases. To better exploit the different importance of the bases, we further proposed a fractional order weighted subspace distance (FOWSD) method and an exponential weighted subspace distance (EWSD) method within the GSD framework in Section 3.4, where discriminative weights were assigned to the bases of each subspace for characterizing their different importance in the similarity measurement. Moreover, in Section 3.5, we also proposed the Affine-FOWSD and

Affine-EWSD to further improve the recognition performance. Owing to the proposed discriminative learning procedure for searching the optimal parameter, the discriminability of the proposed subspace distances was improved and the corresponding recognition performance was enhanced in a series of experiments. Experimental results on two image classification tasks including face recognition and object recognition were presented to show the effectiveness of the proposed scheme in Section 3.6. Next, we will proceed to explore other possible forms of data distribution information to further improve the recognition performance.

Chapter 4

Co-Learned Multi-View Spectral Clustering Approach

4.1 Introduction

In the previous chapter, we discussed the single-model based methods for the task of face recognition with image sets. To benefit from the nice properties such as efficiency in representation and robustness to noises, we investigated the correlation-based methods and proposed subspace distance metrics for measuring the distance between image sets.

Although we have strived to explore the potential of the single-model based methods, their performance seems to encounter a bottleneck, e.g., the best recognition results appear to be stable around 93% on the CMU MoBo database. This phenomenon is probably due to the fact that the single-model based methods, including the linear subspace model and the multivariate Gaussian distribution model, fail to capture the underlying structure of image data with arbitrary nonlinear distributions [12]. In this thesis, we give the explanations/definitions of the image set with "linear distribution" and "nonlinear distribution" respectively as follows:

- An image set with a "linear distribution": Images in this set are distributed in a convex region, thus such two sets can be linearly separated. A Guassian density or a linear subspace can efficiently describe such an image set that corresponds to a convex region.

- An image set with a "nonlinear distribution": Images in this set are distributed in

an arbitrary region (complex and non-convex), thus such two sets cannot be linear separated. A nonlinear manifold is a better representation of such an image set.

It is noted that image sets acquired from real-life environments where facial appearances are of large variations in misalignments, illumination, expression, pose, occlusion, etc., are usually with complex distributions.

Thus more flexible models are required to describe an image set with an intrinsic nonlinear structure. Recently, Wang *et al.* proposed a multi-model based method known as Manifold-Manifold Distance (MMD) [16]. They considered each image set as a nonlinear manifold that is described by using a collection of local models. This corresponds to a method based on multiple models, each of which is described by using a linear subspace. To measure the similarity between two different manifolds, a manifold-to-manifold distance (MMD) was designed that is known as the first attempt to define a distance measure for matching manifolds. Through a series of comparisons with existing single-model based methods, the multi-model based method demonstrated an advantage in describing an image set with intrinsic nonlinear structure, thus resulting in a better recognition performance, as reported in [16]. Motivated by this work, we further explore the multi-model based methods in this chapter in order to better utilize the strong ability of describing an image set.

The rest of this chapter is organized as follows. First, we briefly introduce the MMD method, based on which the motivations of our work are discussed in Section 4.2. Then we propose a Co-learned Multi-view Spectral Clustering (CMSC) method in Section 4.3. The CMSC method integrates multi-view information to divide each image set into multiple clusters. These clusters are more representative and precise to describe a manifold. Based on these improved representations, MMD is applied to match the distances between manifolds. In Section 4.4, the performance of the proposed CMSC method is evaluated. Experimental results show the effectiveness of the proposed CMSC method through comparisons with the state of the arts. Finally, this chapter is summarized in Section 4.5.

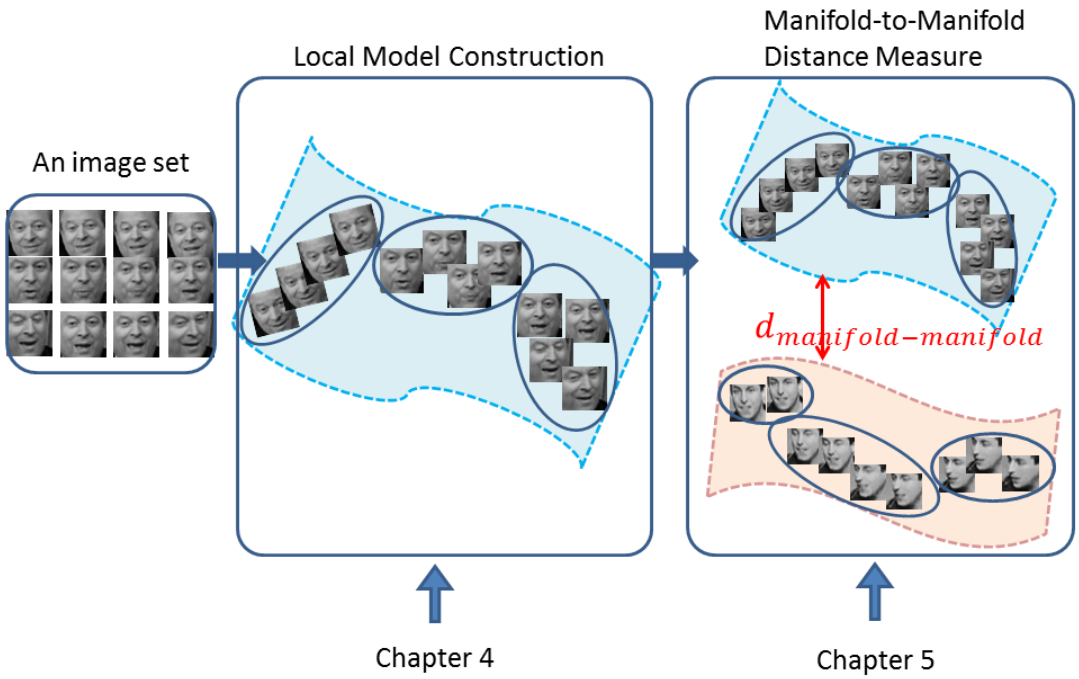

Figure 4.1: The flowchart of the MMD method that consists of two steps: 1. Local model construction; 2. Manifold-to-manifold distance measure. Our work in Chapter 4 focuses on improving the step of local model construction and the enhancement of manifold-to-manifold distance measure will be investigated in Chapter 5.

4.2 Motivations

To better illustrate the motivations and highlight the contributions of our work, we first give a brief introduction on the MMD method [16]. As shown in Fig. 4.1, a flowchart of the MMD method is given, which consists of two steps sequentially, i.e., local model construction and manifold-to-manifold distance measure. In this chapter we will focus on improving the step of local model construction, the enhancement of the manifold-to-manifold distance measure will be investigated in Chapter 5.

Local Model Construction: To facilitate the computation of manifold-to-manifold distance, each image set with nonlinear structure was described by using a collection of local models in the MMD method. A one-shot clustering algorithm was designed that constructs a Maximal Linear Patch (MLP) for each local model such that the linearity property of these local models can be explicitly guaranteed. That is, each local model can be described by using a linear subspace. Then an image set with nearly arbitrary

distributions can be represented by using a collection of linear subspaces.

- **Motivation**:

  (i) A key step of set-based face recognition is how to describe an image set by using a model that exploits the information of this set as much as possible. However, there is a severe under-utilization of image sets in the existing work, where only single view information of each image set is utilized, e.g., the intensity information. In our daily lives, it is common that many real-world datasets naturally consist of multiple views, e.g., multiple different languages of the same article, web pages including contents and hyperlinks, videos including frames and audio information, etc. Even for an image set that only consists of a collection of images, plenty information is still able to be exploited by using some image preprocessing steps, e.g., the SIFT feature [15] and LBP feature [13] [14] can be extracted from an image set and provide complementary information to each other. In view of this, we propose a Co-learned Multi-view Spectral Clustering (CMSC) method to cluster each image set into several clusters by using the information from multiple views simultaneously.

  (ii) We have observed from some preliminary experiments that constructing a linear subspace for a local model is very time-consuming. Furthermore, computing the distance between two manifolds also becomes a costly procedure due to the large computational complexity incurred in the linear subspace matching. In view of these limitations, we adopt a simple but efficient local model, i.e., the mean vector of each local model, to represent image sets. Then a nonlinearly distributed image set can be described by using a collection of mean vectors and the distance between two collections of mean vectors can be easily computed.

- **Contribution**:

  (i) The existing set-based face recognition methods only utilize a single view of a dataset to recognize individuals. In contrast, our proposed CMSC method integrates multiple views of a dataset to enhance the performance.

(ii) Different from the existing multi-view clustering algorithms that divide each image set in an iterative manner [123], the proposed CMSC method solves the optimization function immediately after a relaxation of the constraints. This provides an efficient and robust division of each image set according to its multi-view information.

Manifold-to-Manifold Distance Measure: To compare two collections of linear subspaces, the most efficient way is to compute the distance between the most similar pair of linear subspaces from the two manifolds. This distance was defined as the manifold-to-manifold distance in the MMD method. In Chapter 5, we will proceed to explore more possibilities to calculate the distance between manifolds both efficiently and robustly.

4.3 The Proposed CMSC Method

Among the numerous clustering methods, k-means clustering [124] is one of the most popular algorithms that classify a given data set into a fixed number (the number is pre-defined) of clusters by involving an EM procedure. However, k-means algorithm confronts some limitations, e.g., k-means algorithm cannot produce a satisfactory clustering when the natural clusters do not correspond to convex regions. To solve this problem, we can transform the original data points through an embedding into a low-dimensional subspace where the natural clusters correspond to convex regions and can be linearly separated. This idea was proposed and used in the spectral clustering [125] [126], which clusters the data points based on the first k eigenvectors of a normalized Laplacian matrix of a graph, whose edges denote the similarities between data points.

The Single View Spectral Clustering [126]: Assuming there is a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, from which an adjacency matrix $\mathbf{S}$ that stores the similarities between pairwise points of the data set $\mathbf{X}$ can be computed, i.e., $\mathbf{S}_{i,j} = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma^2}\right)$. Then the Laplacian matrix of this graph can be obtained through $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{S}_{i,j}$ and $\mathbf{D}_{i,j} = 0$ for $i \neq j$. Based on the normalized Laplacian matrix $\mathbf{L}$, the single view spectral clustering algorithm aims to find the optimal embedding matrix denoted as $\mathbf{V}$, which consists of the embedding coordinates of data points in the

low-dimensional subspace:

$$\max_{\mathbf{V}\in\mathbf{R}^{n\times k}} tr\left(\mathbf{V}^T\mathbf{L}\mathbf{V}\right) \qquad s.t.\ \ \mathbf{V}^T\mathbf{V}=\mathbf{I}. \tag{4.1}$$

The optimal $\mathbf{V}$ consists of the first k eigenvectors of the symmetric normalized Laplacian matrix $\mathbf{L}$ as its columns. The i-th row vector of $\mathbf{V}$ provides the embedding coordinates of the i-th data point in the low-dimensional subspace, $i \in \{1, 2, \cdots, n\}$. Thus row vectors of $\mathbf{V}$ can be considered as the relaxations of the indicator vectors and are assigned to one of the k clusters through the k-means algorithm.

The single view spectral clustering has some nice properties such as the well-defined mathematic framework, the more discriminative nature for clustering, good performance on clusters with arbitrary shapes, etc. To exploit these merits, a co-regularized multi-view spectral clustering (Co-regularized MSC) approach was proposed [123] that extended the single view spectral clustering method to the multi-view case through co-training the relaxations of the indicator vectors by using different representations of a data set corresponding to its multiple views. However, some limitations can be observed for this method. For example, the Co-regularized MSC [123] aims to seek a matrix $\mathbf{V}$ for each view of the data set in an iterative manner, which results in only an approximation of the actual solution to the objective function (4.1). Moreover, a kernel matrix was predefined for each matrix $\mathbf{V}$, which was used in the last term of the objective function (4.1) as an additional constraint. Although it facilitates the solving of this optimization, the geometric intuition of the multiplication among these kernel matrices is not clear. Furthermore, although the number of iterations can be pre-defined, the analysis on the convergence of this iterative algorithm was absent.

To overcome the limitations mentioned above, we propose a Co-learned Multi-view Spectral Clustering (CMSC) approach to collaboratively learn the embedding matrix $\mathbf{V}_i$ for each view i of the data set, from which the co-learned embeddings can be used to obtain more accurate clusters. The basic idea of the proposed method is illustrated in Fig. 4.2. Compared with the Co-regularized MSC approach, the proposed CMSC method demonstrates some advantages that are given in the following.

- To expose the geometric intuition explicitly, we use the between-view correlation as a constraint such that the embeddings obtained from different views get closer to

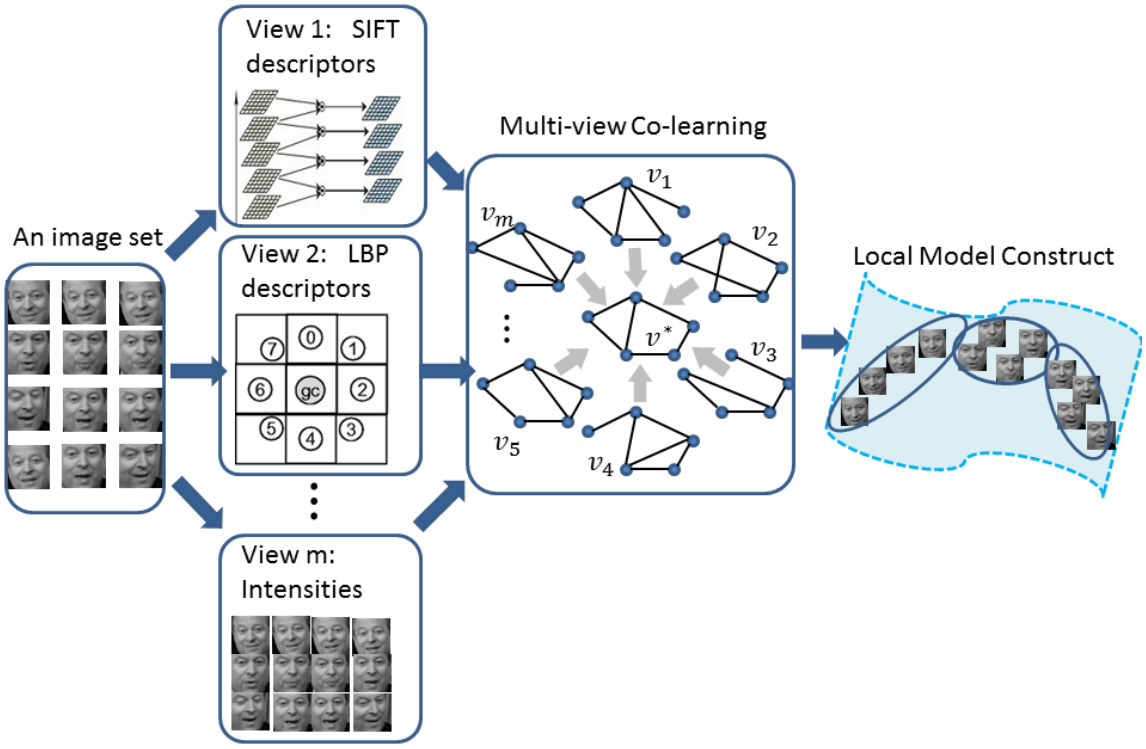

Figure 4.2: The proposed CMSC method clusters an image set into several subsets through a co-learning stage, which enforces the graphs from multiple views to be consistent with each other.

each other. This is consistent with our objective that representations of the same data point in different views should be assigned to the same cluster.

- Instead of an iterative learning procedure, our proposed CMSC method solves the objective function (4.1) through a generalized eigen-decomposition, which leads to a closed-form solution rather than an approximation obtained by Co-regularized MSC. After the generalized eigen-decomposition, the embedding matrices $\mathbf{V}_1, \cdots, \mathbf{V}_m$ for all views of the data set can be obtained simultaneously, which is more efficient and accurate than updating only one matrix in each iteration.

In the following, we first propose two methods for two-view spectral clustering, namely pairwise-based CMSC method and centroid-based CMSC method, and then extend these two methods to the multi-view case respectively.

4.3.1 Pairwise-based CMSC

4.3.1.1 Pairwise-based Two-view Spectral Clustering

Assuming there are n sample points in an image data set. $\mathbf{X}^i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \cdots, \mathbf{x}_n^i]$ and $\mathbf{X}^j = [\mathbf{x}_1^j, \mathbf{x}_2^j, \cdots, \mathbf{x}_n^j]$ denote different representations of this data set under two views i, j. Then the normalized Laplacian matrix $\mathbf{L}_i$ can be obtained through $\mathbf{D}_i^{-1/2}\mathbf{S}_i\mathbf{D}_i^{-1/2}$ and the same applies for $\mathbf{L}_j$.

Since the representation in each single view of the data set is sufficient for clustering independently, we attempt to maximize both $tr(\mathbf{V}_i^T\mathbf{L}_i\mathbf{V}_i)$ and $tr(\mathbf{V}_j^T\mathbf{L}_j\mathbf{V}_j)$ such that the embedding matrix $\mathbf{V}_i$ and $\mathbf{V}_j$, i.e., the approximations of the cluster indicator vectors, can be optimized for view i and view j. Meanwhile, the embeddings obtained from different views of the same data point should be assigned to the same cluster. Thus, the learned embedding matrices $\mathbf{V}_i$ and $\mathbf{V}_j$ should get closer to each other in the low-dimensional subspace. In view of this, we attempt to minimize the distance between the two embedding matrices $\mathbf{V}_i$ and $\mathbf{V}_j$ in our objective function:

$$\begin{aligned}\min_{\mathbf{V}_i,\mathbf{V}_j\in\mathbf{R}^{n\times k}} \|\mathbf{V}_i - \mathbf{V}_j\|_2^2 &= tr\left(\mathbf{V}_i^T\mathbf{V}_i - \mathbf{V}_j^T\mathbf{V}_i - \mathbf{V}_i^T\mathbf{V}_j + \mathbf{V}_j^T\mathbf{V}_j\right) \\ &= 2\cdot k - 2\cdot tr(\mathbf{V}_i^T\mathbf{V}_j),\end{aligned} \tag{4.2}$$

where $tr(\mathbf{V}_i^T\mathbf{V}_i) = tr(\mathbf{V}_j^T\mathbf{V}_j) = k$. Ignoring the constant term $2\cdot k$ and the scaling term 2, the optimization problem in (4.2) can be rewritten as:

$$\max_{\mathbf{V}_i,\mathbf{V}_j\in\mathbf{R}^{n\times k}} tr(\mathbf{V}_i^T\mathbf{V}_j). \tag{4.3}$$

That is, upon maximizing the similarity between two embedding matrices $\mathbf{V}_i$ and $\mathbf{V}_j$ in (4.3), the two graphs get closer to each other in the low-dimensional subspace.

In order to find two embedding matrices $\mathbf{V}_i$ and $\mathbf{V}_j$ that most facilitate the two-view spectral clustering, we attempt to optimize the above three objective functions simultaneously, which can be integrated as below:

$$\begin{aligned}\max_{\mathbf{V}_i,\mathbf{V}_j\in\mathbf{R}^{n\times k}} \lambda_1\Big[tr(\mathbf{V}_i^T\mathbf{L}_i\mathbf{V}_i) + tr(\mathbf{V}_j^T\mathbf{L}_j\mathbf{V}_j)\Big] + \lambda_2\cdot tr(\mathbf{V}_i^T\mathbf{V}_j) \\ s.t. \qquad \mathbf{V}_i^T\mathbf{V}_i = \mathbf{I}, \quad \mathbf{V}_j^T\mathbf{V}_j = \mathbf{I},\end{aligned} \tag{4.4}$$

where parameters λ_1 and λ_2 are introduced to balance the importance between the independent spectral clustering terms and their correlations.

However, both the constrains in (4.4) are nonlinear, which is intractable to obtain closed-form solutions to this optimization problem. Thus we relax them to a single constraint as:

$$s.t. \qquad \beta_i \mathbf{V}_i^T \mathbf{V}_i + \beta_j \mathbf{V}_j^T \mathbf{V}_j = \mathbf{I}, \tag{4.5}$$

where β_i and β_j are introduced to balance the power of constraints in two different views.

For ease of exposition, we let $\mathbf{V} = \begin{bmatrix} \mathbf{V}_i \\ \mathbf{V}_j \end{bmatrix}$, then the objective function in (4.4) can be rewritten in a matrix form as:

$$\begin{aligned} &\max_{\mathbf{V} \in \mathbf{R}^{2n \times k}} tr(\mathbf{V}^T \mathbf{L} \mathbf{V}), \\ &s.t. \qquad \mathbf{V}^T \mathbf{B} \mathbf{V} = \mathbf{I}, \end{aligned} \tag{4.6}$$

where

$$\mathbf{L} = \begin{bmatrix} \lambda_1 \mathbf{L}_i & \lambda_2 \mathbf{I}^{n \times n} \\ 0 & \lambda_1 \mathbf{L}_j \end{bmatrix} \tag{4.7}$$

$$\mathbf{B} = \begin{bmatrix} \beta_i \mathbf{I}^{n \times n} & 0 \\ 0 & \beta_j \mathbf{I}^{n \times n} \end{bmatrix}, \tag{4.8}$$

and $\mathbf{I}^{n \times n}$ denotes an identity matrix with the size of $n \times n$. Then it is observed that the optimization problem in (4.4) becomes a generalized eigen-decomposition of matrices $\mathbf{L}$ and $\mathbf{B}$, i.e.,

$$\mathbf{L}\mathbf{V} = \lambda \mathbf{B}\mathbf{V}. \tag{4.9}$$

4.3.1.2 Pairwise-based Multi-view Spectral Clustering

Next we extend the proposed two-view spectral clustering method to the multi-view case. In the multi-view scenario, we assume that there are representations of a data set in m views, i.e., $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m$, for which the Laplacian matrices $\mathbf{L}_1, \mathbf{L}_2, \cdots, \mathbf{L}_m$ can be calculated respectively. Then the proposed objective function aims to seek the low-dimensional embeddings $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m$ to maximize both the individual spectral clustering terms and their correlations:

$$\begin{aligned} \max_{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m \in \mathbf{R}^{n \times k}} & \lambda_1 \Big[\sum_{i=1}^{m} tr(\mathbf{V}_i^T \mathbf{L}_i \mathbf{V}_i) \Big] + \lambda_2 \Big[\sum_{1 \le i,j \le m, i \ne j} tr(\mathbf{V}_i^T \mathbf{V}_j) \Big] \\ & s.t. \qquad \mathbf{V}_i^T \mathbf{V}_i = \mathbf{I}, \quad \forall i \in \{1, 2, \cdots, m\}, \end{aligned} \tag{4.10}$$

where λ_1 and λ_2 are introduced to balance the importance between the individual spectral clustering terms and their correlations, and $\sum_{1\leq i,j\leq m, i\neq j} tr(\mathbf{V}_i^T\mathbf{V}_j)$ denotes the summation of all the pairwise correlations between graphs in multiple views. To obtain a closed-form solution to this optimization problem, similarly we combine the m nonlinear constraints into a single one as follows:

$$s.t. \qquad \sum_{i=1}^{m} \beta_i \mathbf{V}_i^T \mathbf{V}_i = \mathbf{I}, \tag{4.11}$$

where $\beta_1, \beta_2, \cdots, \beta_m$ are introduced to balance the power of constraints in the m different views.

Similarly, let $\mathbf{V}^T = \begin{bmatrix}\mathbf{V}_1^T & \mathbf{V}_2^T & \cdots & \mathbf{V}_m^T\end{bmatrix}$, the objective function in (4.10) can be rewritten in a matrix form as:

$$\begin{aligned} &\max_{\mathbf{V}\in\mathbf{R}^{(m\cdot n)\times k}} tr(\mathbf{V}^T\mathbf{L}\mathbf{V}), \\ &s.t. \qquad \mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{I}, \end{aligned} \tag{4.12}$$

where

$$\mathbf{L} = \begin{bmatrix} \lambda_1\mathbf{L}_1 & \lambda_2\mathbf{I}^{n\times n} & \cdots & \lambda_2\mathbf{I}^{n\times n} \\ 0 & \lambda_1\mathbf{L}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_2\mathbf{I}^{n\times n} \\ 0 & \cdots & 0 & \lambda_1\mathbf{L}_m \end{bmatrix}, \tag{4.13}$$

$$\mathbf{B} = \begin{bmatrix} \beta_1\mathbf{I}^{n\times n} & 0 & \cdots & 0 \\ 0 & \beta_2\mathbf{I}^{n\times n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \beta_m\mathbf{I}^{n\times n} \end{bmatrix}. \tag{4.14}$$

The constraint $\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{I}$ imposes a scale normalization on this objective function such that a trivial solution can be avoided. In the same way, we can solve the optimization problem in (4.12) by performing a generalized eigen-decomposition $\mathbf{L}\mathbf{V} = \lambda\mathbf{B}\mathbf{V}$.

4.3.2 Centroid-based CMSC

In the pairwise-based CMSC discussed previously, we assume that all the embedding matrices $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m$ tend to each other in the low-dimensional subspace, which requires

computation of the pairwise correlations among m embedding matrices in different views. From (4.10), it is noted that altogether $\frac{m(m-1)}{2}$ multiplications are conducted, which is very time-consuming. In view of this, a centroid-based scenario is considered where we assume that all the embedding matrices $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m$ move towards a centroid matrix $\mathbf{V}^*$ in the low-dimensional subspace. Thus, there are only m comparisons between each embedding matrix $\mathbf{V}_i$, $i \in \{1, \cdots, m\}$, and the centroid matrix $\mathbf{V}^*$ need to be calculated, which significantly reduces the computational complexity of the correlation terms.

In the following, with an extra embedding matrix $\mathbf{V}^*$ called centroid matrix as a reference, we propose a centroid-based CMSC approach that significantly simplifies the computation of the correlation terms. Again, we begin with the case of two-view spectral clustering and then extend it to the multi-view case.

4.3.2.1 Centroid-based Two-view Spectral Clustering

By introducing a centroid matrix $\mathbf{V}^*$, we attempt to seek three embedding matrices $\mathbf{V}_i$, $\mathbf{V}_j$ and $\mathbf{V}^*$ that maximize the objective function as follows:

$$\begin{aligned} \max_{\mathbf{V}_i, \mathbf{V}_j, \mathbf{V}^* \in \mathbf{R}^{n \times k}} \quad & \lambda_1 \Big[tr(\mathbf{V}_i^T \mathbf{L}_i \mathbf{V}_i) + tr(\mathbf{V}_j^T \mathbf{L}_j \mathbf{V}_j) + tr(\mathbf{V}^{*T} \mathbf{L}^* \mathbf{V}^*) \Big] \\ & + \lambda_2 \cdot tr(\mathbf{V}_i^T \mathbf{V}^* + \mathbf{V}_j^T \mathbf{V}^*) \\ & s.t. \quad \mathbf{V}_i^T \mathbf{V}_i = \mathbf{I}, \quad \mathbf{V}_j^T \mathbf{V}_j = \mathbf{I}, \quad \mathbf{V}^{*T} \mathbf{V}^* = \mathbf{I}, \end{aligned} \tag{4.15}$$

where λ_1 and λ_2 are introduced to balance the importance between the individual spectral clustering terms and the correlation terms. Similar to the pairwise-based case, to obtain a closed-form solution, we relax the three nonlinear constraints by combining them into a single one:

$$s.t. \qquad \beta_i \mathbf{V}_i^T \mathbf{V}_i + \beta_j \mathbf{V}_j^T \mathbf{V}_j + \beta^* \mathbf{V}^{*T} \mathbf{V}^* = \mathbf{I}, \tag{4.16}$$

where β_i, β_j and β^* are pre-defined according to the prior knowledge and are used to balance the power of constraints in two different views. This constraint imposes a scale normalization on the objective function such that a trivial solution can be avoided.

Similarly, to facilitate the solving of the objective function, we let $\mathbf{V}^T = [\mathbf{V}_i^T \ \mathbf{V}_j^T \ \mathbf{V}^{*T}]$, then the optimization problem restricted by the relaxed constraint in (4.16) can be rewritten in a matrix form as:

$$\begin{aligned} & \max_{\mathbf{V} \in \mathbf{R}^{3n \times k}} tr(\mathbf{V}^T \mathbf{L} \mathbf{V}), \\ & s.t. \qquad \mathbf{V}^T \mathbf{B} \mathbf{V} = \mathbf{I}, \end{aligned} \tag{4.17}$$

where

$$\mathbf{L} = \begin{bmatrix} \lambda_1\mathbf{L}_i & 0 & \lambda_2\mathbf{I}^{n\times n} \\ 0 & \lambda_1\mathbf{L}_j & \lambda_2\mathbf{I}^{n\times n} \\ 0 & 0 & \lambda_1\mathbf{L}^* \end{bmatrix}, \tag{4.18}$$

$$\mathbf{B} = \begin{bmatrix} \beta_i\mathbf{I}^{n\times n} & 0 & 0 \\ 0 & \beta_j\mathbf{I}^{n\times n} & 0 \\ 0 & 0 & \beta^*\mathbf{I}^{n\times n} \end{bmatrix}, \tag{4.19}$$

and $\mathbf{L}^*$ is the Laplacian matrix of the centroid graph $\mathbf{S}^*$. To reasonably initialize the centroid graph $\mathbf{S}^*$, we let $\mathbf{S}^* = \frac{\mathbf{S}_i+\mathbf{S}_j}{2}$, then we have the Laplacian matrix $\mathbf{L}^* = \mathbf{D}^{*-1/2}\mathbf{S}^*\mathbf{D}^{*-1/2}$.

Similarly, this optimization problem becomes a generalized eigen-decomposition of matrices $\mathbf{L}$ and $\mathbf{B}$, i.e., $\mathbf{LV} = \lambda\mathbf{BV}$.

4.3.2.2 Centroid-based Multi-view Spectral Clustering

Next we extend the proposed centroid-based two-view spectral clustering method to the multi-view case. In the multi-view scenario, the proposed objective function aims to seek $m+1$ low-dimensional embeddings $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m, \mathbf{V}^*$ to maximize both the individual spectral clustering terms and their correlations:

$$\begin{aligned} \max_{\mathbf{V}_1,\mathbf{V}_2,\cdots,\mathbf{V}_m,\mathbf{V}^*\in\mathbf{R}^{n\times k}} & \lambda_1\Big[\sum_{i=1}^{m} tr(\mathbf{V}_i^T\mathbf{L}_i\mathbf{V}_i) + tr(\mathbf{V}^{*T}\mathbf{L}^*\mathbf{V}^*)\Big] + \lambda_2\Big[\sum_{i=1}^{m} tr(\mathbf{V}_i^T\mathbf{V}^*)\Big] \\ s.t. \quad & \mathbf{V}_i^T\mathbf{V}_i = \mathbf{I}, \quad \forall i \in \{1, 2, \cdots, m\}, \quad \mathbf{V}^{*T}\mathbf{V}^* = \mathbf{I}, \end{aligned} \tag{4.20}$$

where λ_1, λ_2 are introduced to balance the importance between the individual spectral clustering terms and their correlations, and $\sum_{i=1}^{m} tr(\mathbf{V}_i^T\mathbf{V}^*)$ denotes the summation of all the correlations that are computed between each embedding matrix $\mathbf{V}_i$, $i \in \{1, 2, \cdots, m\}$, and the centroid matrix $\mathbf{V}^*$.

Similarly, to obtain a closed-form solution, the multiple nonlinear constraints in (4.20) can be combined into a single one:

$$s.t. \quad \sum_{i=1}^{m} \beta_i\mathbf{V}_i^T\mathbf{V}_i + \beta^*\mathbf{V}^{*T}\mathbf{V}^* = \mathbf{I}, \tag{4.21}$$

where $\beta_1, \beta_2, \cdots, \beta_m$ and β^* are used to balance the power of constraints in the m different views and the centroid embeddings. Again, we let $\mathbf{V}^T = \Big[\mathbf{V}_1^T \ \mathbf{V}_2^T \ \cdots \ \mathbf{V}_m^T \ \mathbf{V}^{*T}\Big]$,

then the objective function in (4.20) restricted by the relaxed constraint in (4.21) can be rewritten in a matrix form as:

$$\begin{aligned} &\max_{\mathbf{V}\in\mathbf{R}^{(m+1)\cdot n\times k}} tr(\mathbf{V}^T\mathbf{L}\mathbf{V}), \\ &s.t. \quad \mathbf{V}^T\mathbf{B}\mathbf{V}=\mathbf{I}, \end{aligned} \tag{4.22}$$

where

$$\mathbf{L} = \begin{bmatrix} \lambda_1\mathbf{L}_1 & 0 & \cdots & 0 & \lambda_2\mathbf{I}^{n\times n} \\ 0 & \lambda_1\mathbf{L}_2 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \lambda_1\mathbf{L}_m & \lambda_2\mathbf{I}^{n\times n} \\ 0 & \cdots & \cdots & 0 & \lambda_1\mathbf{L}^* \end{bmatrix}, \tag{4.23}$$

$$\mathbf{B} = \begin{bmatrix} \beta_1\mathbf{I}^{n\times n} & 0 & \cdots & \cdots & 0 \\ 0 & \beta_2\mathbf{I}^{n\times n} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta_m\mathbf{I}^{n\times n} & 0 \\ 0 & \cdots & \cdots & 0 & \beta^*\mathbf{I}^{n\times n} \end{bmatrix}, \tag{4.24}$$

$\mathbf{L}^* = \mathbf{D}^{*-1/2}\mathbf{S}^*\mathbf{D}^{*-1/2}$, and $\mathbf{S}^* = \frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i$.

Similarly, the optimization problem becomes a generalized eigen-decomposition where $\mathbf{L}\mathbf{V} = \lambda\mathbf{B}\mathbf{V}$. And by imposing the constraint $\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{I}$, we avoid a trivial solution to the objective function.

4.3.3 Discussions

From the above analysis, there are some issues that we should pay attention to for the proposed approaches to work properly.

- It is noticed that the large Laplacian matrices obtained in (4.7), (4.13), (4.18) and (4.23), i.e., $\mathbf{L}$, are not symmetric matrices. Then the eigenvectors that are calculated through the generalized eigen-decomposition of matrices $\mathbf{L}$ and $\mathbf{B}$ are only linear independent but not guaranteed to be orthogonal to each other. To deal with this, we apply symmetrization on the large Laplacian matrix $\mathbf{L}$ for obtaining a set of orthogonal eigenvectors:

$$\mathbf{L} = (\mathbf{L}^T + \mathbf{L})/2. \tag{4.25}$$

- After obtaining the m low-dimensional embeddings $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_m$, we normalize each row of the embedding matrix $\mathbf{V}_i$, $i \in \{1, \cdots, m\}$. Then a k-means clustering algorithm is applied to each normalized matrix $\mathbf{V}_i$ to accomplish the assignments, i.e., the p^{th} data point is assigned to one of the clusters according to the clustering result of the p^{th} row of the embedding matrix $\mathbf{V}_i$. In this way, a image set is clustered into several subsets, each of which is represented by a local model. Thus, a nonlinearly distributed manifold can be described by using a collection of local models in our proposed method.

- When the number of views $m \leq 3$, the computational complexity of the pairwise-based CMSC method is equivalent or slightly lower than that of the centroid-based CMSC method. To be specific, it is suggested to use the pairwise-based CMSC method when $m = 2$. However, when the number of views $m > 3$, the computational complexity of the pairwise-based CMSC method becomes much higher than that of the centroid-based CMSC approach. In this case, the centroid-based CMSC approach is preferred.

- Once the matrix $\mathbf{V}$ is obtained, the embedding matrix $\mathbf{V}_1$ can be derived from the first n rows of $\mathbf{V}$, and the embedding matrix $\mathbf{V}_2$ can be obtained from the second n rows of $\mathbf{V}$, so on and so forth.

4.3.4 The Whole Procedure

In this section, we describe the whole procedure of the proposed CMSC approaches for face recognition with image sets.

(i) **Image Set Modeling**: Assuming there are N face image sets stored in the gallery dataset. For each image set, it is supposed that the face images (data points) lie on an underlying nonlinear low dimensional manifold $\mathbf{M}$. For the image sets in the gallery dataset, we apply the proposed CMSC approaches to exploit the information in multiple views of data sets such that more accurate clusters can be obtained for each image set. In this way, a nonlinear manifold $\mathbf{M}$ can be divided into multiple subsets (clusters) and each subset is described by its mean vector $\mathbf{m}_i$, i.e., a mean

vector denotes a local model. Taking into account the fact that the number of samples in each image set is usually not sufficient enough to be precisely modeled by using a Gaussian distribution, a linear subspace or an affine hull, we adopt the mean vector of the limited samples as a more appropriate and representative descriptor to describe each subset. In this way, a manifold $\mathbf{M}$ can be modeled by using a collection of mean vectors $\{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_r\}$.

(ii) **Manifolds Matching**: As pointed out in [104] [16], the most effective way to measure the similarity between two manifolds is to compute the distance between their most similar parts. Inspired by this conclusion, we first model each manifold as a collection of mean vectors, then we compute the nearest distance between local models from different manifolds as the distance measure between the two manifolds $\mathbf{M}^X = \left\{\mathbf{m}_1^X, \mathbf{m}_2^X, \cdots, \mathbf{m}_{n_X}^X\right\}$ and $\mathbf{M}^Y = \left\{\mathbf{m}_1^Y, \mathbf{m}_2^Y, \cdots, \mathbf{m}_{n_Y}^Y\right\}$, which is given as

$$d(\mathbf{M}^X, \mathbf{M}^Y) = \min\left\{\min_i d(\mathbf{m}_i^X, \mathbf{M}^Y), \min_j d(\mathbf{m}_j^Y, \mathbf{M}^X)\right\}, \tag{4.26}$$

where

$$d(\mathbf{m}_i^X, \mathbf{M}^Y) = \min_{1\le j\le n_Y} d(\mathbf{m}_i^X, \mathbf{m}_j^Y) = \|\mathbf{m}_i^X - \mathbf{m}_j^Y\|_2^2, \tag{4.27}$$

$$d(\mathbf{m}_j^Y, \mathbf{M}^X) = \min_{1\le i\le n_X} d(\mathbf{m}_i^X, \mathbf{m}_j^Y) = \|\mathbf{m}_i^X - \mathbf{m}_j^Y\|_2^2, \tag{4.28}$$

or

$$d(\mathbf{m}_i^X, \mathbf{M}^Y) = \min_{1\le j\le n_Y} d(\mathbf{m}_i^X, \mathbf{m}_j^Y) = \frac{(\mathbf{m}_i^X)^T\mathbf{m}_j^Y}{\|\mathbf{m}_i^X\|_2\|\mathbf{m}_j^Y\|_2}, \tag{4.29}$$

$$d(\mathbf{m}_j^Y, \mathbf{M}^X) = \min_{1\le i\le n_X} d(\mathbf{m}_i^X, \mathbf{m}_j^Y) = \frac{(\mathbf{m}_i^X)^T\mathbf{m}_j^Y}{\|\mathbf{m}_i^X\|_2\|\mathbf{m}_j^Y\|_2}, \tag{4.30}$$

alternatively, where $i \in \{1, 2, \cdots, n_X\}$, $j \in \{1, 2, \cdots, n_Y\}$, $d(\mathbf{m}_i^X, \mathbf{M}^Y)$ denotes the distance between local model $\mathbf{m}_i^X$ and manifold $\mathbf{M}^Y$, and $d(\mathbf{m}_i^X, \mathbf{m}_j^Y)$ denotes the pairwise distance between the local models $\mathbf{m}_i^X$ and $\mathbf{m}_j^Y$.

(iii) **Recognition**: Given a gallery set containing N face image sets, each image set is represented by a collection of local models as described above. Then a probe set $\mathbf{M}^P$ is compared with every image set $\mathbf{M}_g^G$ in the gallery set and assigned to the class g^* with the highest similarity score

$$g^* = \arg\min_{g\in G} d(\mathbf{M}_g^G, \mathbf{M}^P). \tag{4.31}$$

4.4 Experiments

In this section, we evaluate the two proposed approaches, namely the pairwise-based co-learned multi-view spectral clustering method (PW-CMSC) and the centroid-based co-learned multi-view spectral clustering method (CT-CMSC), on two widely used face video databases, i.e., the Honda/UCSD database [127] and the Youtube Celebrities [128] database.

4.4.1 Experimental Settings

In the experiments, we aim to evaluate the performance of the proposed PW-CMSC and CT-CMSC under some system configurations that are introduced in the following.

1. Local Models: As introduced in Section 4.2, the multi-model based approaches often consist of two steps: 1. local model construction; 2. manifold-to-manifold distance measure. In our experiments, to better illustrate the effectiveness of the proposed CMSC method, we use 5 different methods to model each face image set.

- **MLP**:

  Maximal Linear Patch (MLP) is a local linear model that was proposed in the reference [16]. In this method, each nonlinear manifold is represented by using a collection of MLPs. Each MLP originates from a seed data point that utilizes the neighboring points to increase itself according to a linear constraint. Then the increased set of data points can be used to span a linear subspace where the linearity of the subspace is controlled by the ratio between Euclidean distance and geodesic distance of the two neighbor points. The points that span the linear subspace keep increasing until the linearity of the subspace is broken. Then we can obtain the maximal linear subspace as the MLP.

- **k-means Clustering**:

  k-means clustering [124] is one of the most popular and the simplest learning algorithms that classifies a given data set into a fixed number (the number is predefined) of clusters by involving an EM procedure. However, k-means algorithm also faces some limitations such as unspecified initialization, dependence on the

value of k, convergence into local minima, etc. In view of these limitations, the value of k can be adaptively set according to the number of face images in each data set. Preliminary experiments demonstrate that with the number of face images in a data set varying from 8 to 1000, k takes values in range of $[3, 29]$.

- **Single View Spectral Clustering**:

  The standard spectral clustering approach was proposed in [125] [126], where the data points are clustered into several subsets by utilizing the single view representation of the data set. In our experiments, the number of eigenvectors obtained is set to be the same or 1.5 times of the number of clusters. In the experiments, this single view spectral clustering method is denoted as "Single-view-SC" for simplicity.

- **Co-regularized Multi-view Spectral Clustering**:

  This method was proposed in reference [123] recently. By using different representations of a data set corresponding to its multiple views, this method extended the single view spectral clustering method to the multi-view case through co-training the relaxations of the indicator vectors. In our experiments, we consider the same configurations where the number of iterations for solving the optimization problem is set around 5 and tuned to the optimal value according to the recognition results. For simplicity, this Co-regularized multi-view spectral clustering method is denoted as "Co-regularized-MS".

- **Proposed PW-CMSC and CT-CMSC**:

  For ease of analysis, for the two proposed CMSC methods, we enforce $\beta_1 + \beta_2 + \cdots + \beta_m = 1$, and set the number of eigenvectors of each embedding matrix $\mathbf{V}_i$ to be 40.

2. Manifold-to-manifold Distance Measure: For fair comparisons, we utilize the same manifold-to-manifold distance measure for all the approaches mentioned above, through which the efficiency of the proposed PW-CMSC and CT-CMSC methods is illustrated in modeling an image set. Since it is very efficient to compute the distance only between the most similar parts of two differen manifolds, we use the manifold-to-manifold distance defined in (4.26), where the distance between the nearest pair of

local models from two different manifolds is used as the dissimilarity between the two manifolds.

3. Databases: In our experiments, we evaluate all the approaches mentioned above on two widely used video face databases, where each video sequence is saved as an image set.

- **The Honda/UCSD Database**:

  The Honda/UCSD database [127] was built for evaluating algorithms on the video based face tracking and face recognition tasks. All the video sequences in this database were captured in indoor environment where the noises and the variations in lighting conditions are limited. However, the person in each video rotates his/her head and makes different expressions, which may lead to large variations in pose, expression, occlusion and scaling. In the Honda/UCSD database, there are 59 video sequences from 20 persons. The length of each video sequence is around 15 seconds. The original resolution of the frame of each video sequence is 640×480 pixels.

- **The Youtube Celebrities Database**:

  The YouTube Celebrities database [128] was built for evaluating algorithms on the tasks of face tracking and recognition. 1910 video sequences were collected from 47 celebrities from the Youtube website. These videos were captured from famous persons including actors/actresses and politicians. Compared to Honda/UCSD and CMU MoBo databases where the videos are captured under controlled conditions, the videos in Youtube Celebrities database are obtained in real life environments where there are very large variations in face appearances, poses and expressions, together with low quality, misalignments and occlusions. The size of each frame of a video sequence varies from 180×240 pixels to 240×320 pixels.

4. Multiple Views: In our experiments, we use both the intensity value and the local LBP features to describe an image set, where each image set is clustered by applying the proposed PW-CMSC, CT-CMSC, and the Co-regularized multi-view spectral clustering approaches. For each face image, we first divide it into multiple 50% overlapped local patches. The size of each local patch is 10×10 pixels for the Honda/UCSD database

Figure 4.3: Some examples of the cropped face images from the Honda/UCSD database.

and is 20×20 pixels for the Youtube Celebrities Database, respectively. Then we extract a LBP descriptor from each local patch and concatenate them into a high-dimensional feature vector to represent the whole face image.

4.4.2 Experimental Results on the Honda/UCSD Database

In this experiment, each video sequence is fragmented into a set of frames, based on which we apply the Viola-Jones face detection algorithm [47] to detect face regions. Then each face image is cropped to 20×20 pixels to facilitate the computation as well as reduce the memory requirement. Some examples of cropped face image sets in the Honda/UCSD database are shown in Fig. 4.3.

It is noted that the clustering algorithms only produce the cluster label for each data point. Based on the label information, the local model construction can be accomplished under two situations: based on the intensity images or based on the LBP descriptors. For both situations, we evaluate different methods under the same configurations as considered in references [117] [98] [127] where 20 video sequences are selected from 20 different subjects as the gallery image sets, with the remaining 39 video sequences used for testing.

When there are large variations in pose and occlusion in the video, the face tracking algorithms may lose the face target, which results in non-continuous face images tracked in real-life applications. In view of this, we evaluate the performance of different methods in three scenarios respectively: 1. the first 50 frames of each video sequence are used; 2. the first 100 frames of each video sequence are used; 3. the full length of a video sequence are used. The experimental results based on the intensity images are shown in Table 4.1 and Table 4.2 respectively for these three scenarios. The average results of the three scenarios are also presented in Table 4.2. Similarly, the experimental results based on the LBP descriptors are shown in Table 4.3 and Table 4.4 respectively, with the average results of the three scenarios given in Table 4.4.

Table 4.1: Face recognition results of the six comparable methods on the intensity images of the Honda/UCSD database, where 50 and 100 frames are selected from each video respectively.

Methods	50 frames	100 frames
MLP [16]	$82.05\% \pm 0.00\%$	$82.05\% \pm 0.00\%$
K-means [124]	$83.85\% \pm 2.14\%$	$91.11\% \pm 2.91\%$
Single-view-SC [126]	$84.62\% \pm 2.56\%$	$93.59\% \pm 1.48\%$
Co-regularized-MSC [123]	$84.62\% \pm 1.81\%$	$94.02\% \pm 1.81\%$
Proposed PW-CMSC	$85.64\% \pm 2.19\%$	$\mathbf{94.87}\% \pm 1.48\%$
Proposed CT-CMSC	$\mathbf{86.15}\% \pm 2.29\%$	$94.36\% \pm 1.24\%$

Table 4.2: Face recognition results of the six comparable methods on the intensity images of the Honda/UCSD database where all frames of each video are used, together with the average results of all three scenarios.

Methods	Full length	Averages
MLP [16]	$94.87\% \pm 0.00\%$	$86.32\% \pm 0.00\%$
K-means [124]	$92.74\% \pm 2.53\%$	$89.24\% \pm 2.53\%$
Single-view-SC [126]	$92.74\% \pm 2.16\%$	$90.31\% \pm 2.07\%$
Co-regularized-MSC [123]	$94.02\% \pm 3.02\%$	$90.89\% \pm 1.61\%$
Proposed PW-CMSC	$95.56\% \pm 2.42\%$	$92.02\% \pm 1.54\%$
Proposed CT-CMSC	$\mathbf{97.44}\% \pm 2.81\%$	$\mathbf{92.65}\% \pm 2.11\%$

From Table 4.1-Table 4.4, some observations can be obtained as follows. (1) In terms of the average results over all three scenarios, the proposed CT-CMSC method

Table 4.3: Face recognition results of the six comparable methods on the LBP descriptors of the Honda/UCSD database, where 50 and 100 frames are selected from each video respectively.

Methods	50 frames	100 frames
MLP [16]	$84.62\% \pm 0.00\%$	$92.31\% \pm 0.00\%$
K-means [124]	$86.32\% \pm 5.33\%$	$90.60\% \pm 2.96\%$
Single-view-SC [126]	$89.91\% \pm 2.23\%$	$97.44\% \pm 3.63\%$
Co-regularized-MSC [123]	$89.74\% \pm 2.56\%$	$97.44\% \pm 2.56\%$
Proposed PW-CMSC	$89.23\% \pm 1.15\%$	$\mathbf{97.95}\% \pm 1.14\%$
Proposed CT-CMSC	$\mathbf{90.26}\% \pm 2.15\%$	$96.92\% \pm 1.14\%$

Table 4.4: Face recognition results of the six comparable methods on the LBP descriptors of the Honda/UCSD database where all frames of each video are used, together with the average results of all three scenarios.

Methods	Full length	Averages
MLP [16]	$94.87\% \pm 0.00\%$	$90.60\% \pm 0.00\%$
K-means [124]	$94.02\% \pm 1.48\%$	$90.33\% \pm 3.26\%$
Single-view-SC [126]	$96.50\% \pm 2.82\%$	$94.62\% \pm 2.89\%$
Co-regularized-MSC [123]	$97.44\% \pm 2.56\%$	$94.87\% \pm 2.56\%$
Proposed PW-CMSC	$98.46\% \pm 2.29\%$	$95.21\% \pm 1.53\%$
Proposed CT-CMSC	$\mathbf{99.48}\% \pm 1.15\%$	$\mathbf{95.55}\% \pm 1.48\%$

outperforms all the other methods, which reaches the average recognition rate of 92.65% and 95.55% based on the intensity images and the LBP descriptors respectively. (2) It is observed that the recognition performance of the CT-CMSC method is better than that of the PW-CMSC method. A possible reason is that the CT-CMSC method applies the k-means clustering on the centroid embedding matrix $\mathbf{V}^*$, rather than the concatenated $\mathbf{V}^T = [\mathbf{V}_1^T \mathbf{V}_2^T \cdots \mathbf{V}_m^T]$ used in the PW-CMSC method. (3) When the number of face images selected from each image set is insufficient (50 or 100), the proposed methods demonstrate better performance than the others. This implies that the proposed methods are able to better process the noisy videos in real-life applications where some videos are not continuous or with very short length. This is because when the number of samples in each image set is limited, it is more suitable to use the mean vector to describe a subset than a complex model such as a Gaussian density or a linear subspace. (4) It is noted that there is no variance for the performance of the method using MLP model. Different from the k-means based approaches that start the clustering from k randomly selected

Figure 4.4: Some examples of the cropped face images from the Youtube Celebrities database.

points, a MLP model only starts from one seed point to construct local models. This makes the MLP much more robust than the k-means based approaches. Whereas the other 5 clustering methods, which are based on the k-means clustering algorithm, cannot achieve a recognition performance with zero-variance.

4.4.3 Experimental Results on the Youtube Celebrities Database

Due to the low quality of the Youtube Celebrities Database, the Viola-Jones face detection algorithm [47], which has strong ability to detect frontal faces, often fails to detect the face regions of the noisy videos in this database. Fortunately, the initials that contain the position information of the face region in the first frame of each video are provided together with the Youtube Celebrities Database. Then we can apply the incremental visual tracking algorithm [129] to track face regions across frames of each video by using the initials that are manually labeled. As shown in Fig. 4.4, some examples of the well-tracked faces are given where the tracked face region of each frame is cropped and resized to 50×50 pixels.

Similarly, we conduct the experiments on the Youtube Celebrities Database under

Table 4.5: Averaged results of the six comparable methods on the intensity images of the Youtube Celebrities database.

Methods	Recognition Rates
MLP [16]	$54.78\% \pm 2.16\%$
K-means [124]	$55.52\% \pm 2.33\%$
Single-view-SC [126]	$55.32\% \pm 1.25\%$
Co-regularized-MSC [123]	$55.32\% \pm 1.25\%$
Proposed PW-CMSC	$\mathbf{56.74}\% \pm 0.50\%$
Proposed CT-CMSC	$\mathbf{56.74}\% \pm 0.50\%$

Table 4.6: Averaged results of the six comparable methods on the LBP descriptors of the Youtube Celebrities database.

Methods	Recognition Rates
MLP [16]	$60.41\% \pm 2.16\%$
K-means [124]	$60.11\% \pm 0.75\%$
Single-view-SC [126]	$60.99\% \pm 1.26\%$
Co-regularized-MSC [123]	$61.35\% \pm 2.02\%$
Proposed PW-CMSC	$62.77\% \pm 1.01\%$
Proposed CT-CMSC	$\mathbf{63.12}\% \pm 0.50\%$

two situations, i.e., based on the intensity images and based on the LBP descriptors respectively. For both situations, we consider the same configurations as adopted in reference [117]. We conduct 5-fold cross validation experiments, where the whole database is divided into 5 folds and each of them contains 423 videos from 47 persons with 9 videos per person. In each fold, 3 image sets per person are randomly selected to constitute the gallery set, with the remaining 6 videos per person used as the probe sets. We repeat the random division 50 times, from which the averaged recognition rates based on the intensity images and based on the LBP descriptors are summarized in Table 4.5 and Table 4.6 respectively.

From Table 4.5 and 4.6, some observations can be obtained as follows. (1) Both the proposed PW-CMSC and CT-CMSC methods achieve the best recognition rate of 56.74% based on the intensity images, which is averaged over all 5 folds and outperforms all the other methods. The proposed CT-CMSC method achieves the best recognition rate of 63.12% based on the LBP descriptors and outperforms all the other methods by around 3%. (2) It is noted that the performance improvements of the proposed methods

on the Honda/UCSD database is larger than that on the Youtube Celebrities database. A possible reason is that the images in Honda/UCSD database are of good qualities such that the intensity values and the LBP descriptors are able to describe the database more completely. In contrast, the images in the Youtube Celebrities database are usually obtained in real-life environment that contain severe noises. Thus both intensity and LBP information is not sufficient to capture enough information for recognition. (3) Since the results shown in Table 4.5 and 4.6 are obtained by averaging the recognition rates of 5 folds where the training data and testing data are randomly selected, the recognition performance of the MLP method is no longer of zero-variance.

4.4.4 Computational Complexity

To illustrate the computational efficiency of the proposed CMSC method, we have done some comparisons between the computational complexities of the CMSC and Co- regularized MSC. To illustrate the timing or time consumption of the algorithms, we have conducted new experiments on a workstation equipped with Intel(R) Xeon(TM) MV CPU 3.20 GHz (2 processors) and 20.0 GB RAM. On the Honda/UCSD database, the computation times of the Co-regularized MSC is 15 minutes and 37 seconds, the proposed CT-CMSC and the proposed PW-CMSC is 4 minutes and 13 seconds, and 2 minutes and 19 seconds, respectively. Thus, our proposed algorithms (PW-CMSC and CT-CMSC) are more efficient than the iterative algorithm (Co-regularized MSC) in terms of timing or computational complexity.

4.5 Conclusion

In this chapter, we investigated the multi-model based approaches on face recognition with image sets. In Section 4.1 and 4.2, we introduced the framework of the multi-model based approaches that consists of two procedures, i.e., local model construction and manifold-to-manifold distance measure. In this chapter, we attempted to improve the procedure of local model construction. A pairwise-based multi-view spectral clustering approach and a centroid-based multi-view spectral clustering approach were proposed in Section 4.3 to utilize the information provided by multiple views of each image set. Based

on the improved clusters, the mean vectors become more representative, which leads to improvements on the recognition performance of the proposed methods, as verified by the experimental results in Section 4.4. As a continuation, we will proceed to investigate the multi-model based approaches in the next chapter, where the manifold-to-manifold distance measure will be further studied.

Chapter 5

Collaborative Reconstruction-Based Manifold-to-Manifold Distance

5.1 Introduction

In the previous chapter, we began to discuss the multi-model based methods for matching image sets with nonlinear distributions and investigated the stage of local model construction. In this chapter, we will focus on the stage of manifold-to-manifold distance measure, as a continuous of the discussion in Chapter 4.

In the face recognition based on image sets scenario, the task of comparing the gallery image sets and a probe image set becomes the problem of defining an appropriate manifold-to-manifold distance, as demonstrated in Fig. 5.1. The first attempt to explicitly define the manifold-to-manifold distance is the method called MMD proposed in reference [16]. In their method, the authors considered that the most efficient way to measure the similarity of two manifolds is finding the most similar parts of the two manifolds. Thus, the pairwise distances between local models from two different manifolds are computed and the distance with the smallest value is selected and considered as the dissimilarity of the two manifolds. For easy of explanation, we denote the manifold-to-manifold distance described above as the Nearest Neighbor-based MMD (NN-MMD). Obviously, NN-MMD is a straightforward design of the manifold-to-manifold distance and has several limitations. For example, it cannot guarantee a reliable recognition result when an face image set contains outliers, since the minimal pairwise distance of local models may be computed based on the outliers of the two manifolds.

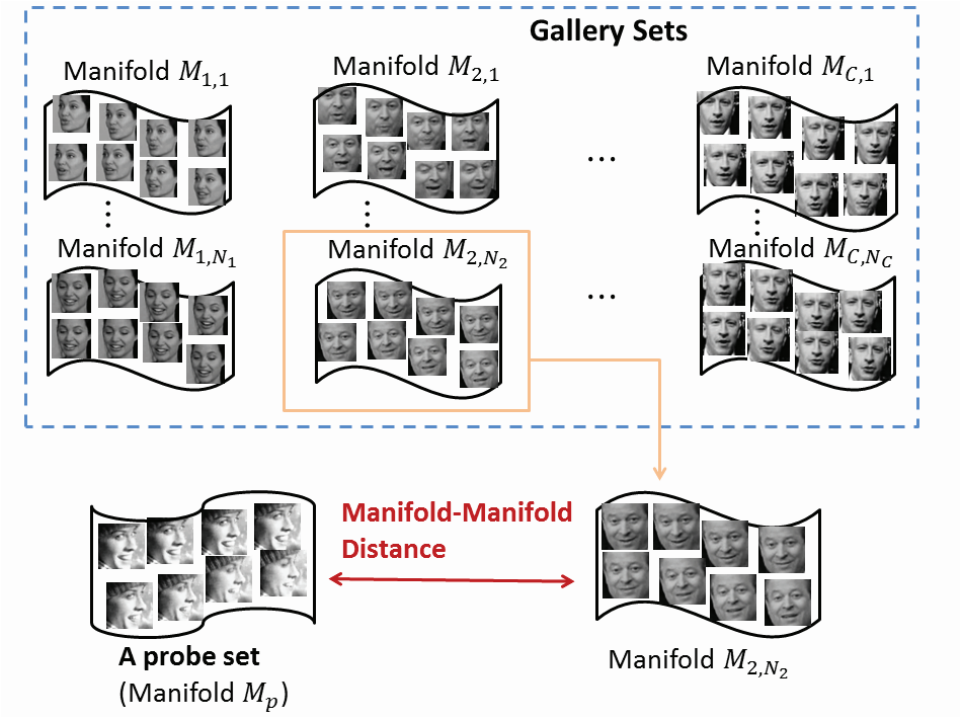

Figure 5.1: For face recognition based on image sets scenario, the problem of comparing the gallery image sets and a probe image set becomes the problem of defining an appropriate manifold-to-manifold distance.

To handle this problem, we propose a robust and flexible manifold-to-manifold distance measure called Collaborative Reconstruction-Based Manifold-to-Manifold Distance (CRMMD) for face recognition based on image sets. In our proposed method, we first divide each manifold (one image set) into multiple clusters through a clustering approach that could be the k-means clustering algorithm [130] or one of the CMSC methods proposed in the previous chapter. Then a manifold is described by a collection of mean vectors. Based on these mean vectors, designing the manifold-to-manifold distance becomes easier than based on other complex local models. Instead of computing the pairwise distance between local models, we calculate the distance between a local model and its approximation, which is collaboratively reconstructed from several local models on other manifolds. Hence, the proposed CRMMD reduces its sensitivity to random variations in local models and provides more accurate classification results.

The rest of this chapter is organized as follows: The proposed Collaborative Reconstruction Based Manifold-to-Manifold Distance (CRMMD) will be discussed in Section

5.2, which focuses on the second stage of the multi-model based methods, i.e., measuring the similarity between different manifolds. Section 5.3 evaluates the proposed CRMMD method and compared it with several popular methods that are based on unsupervised image set matching. Experimental results shows the efficacy of the proposed CRMMD method. Section 5.4 conclusions our investigation on the two stages of the multi-model based methods in this chapter.

5.2 Proposed CRMMD

Assuming there are N image sets in the gallery dataset. For each set, we consider that face samples lie on an underlying nonlinear and intrinsic low dimensional manifold $\mathbf{M}$. We first divide a nonlinear manifold $\mathbf{M}$ into several subsets by adopt a certain clustering approach that can be the k-means clustering algorithm [130] or one of the CMSC methods proposed in the previous chapter. These clustering approaches are able to provide a stable and representative mean vector for each cluster such that we use the mean vector $\mathbf{m}_i$ to describe each subset i. In this way, a mean vector denotes a local model. There are several advantages of using mean vectors to describe local models. For example, when the number of faec images in each subset is very limited, we cannot precisely model them using Gaussian distribution, linear subspace or affine hull, thus the mean vector of the limited face images becomes a more appropriate and representative descriptor adopted to describe each subset. Furthermore, when a manifold is described by using a collection of mean vectors, the distance between two collections of mean vectors can be computed rapidly and flexibly as the manifold-to-manifold distance. Moreover, it provides us more possibilities to design an appropriate manifold-to-manifold distance based on the collections of mean vectors. Thus, a collection of mean vectors $\{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_r\}$ is used in our work to model the manifold $\mathbf{M}$.

As discussed in Section 5.1, the existing manifold-to-manifold distance measure (NN-MMD) may be sensitive to outliers that are often obtained under the unconstrained conditions. To solve this problem, we propose a robust and flexible manifold-manifold distance measure called Collaborative Reconstruction-Based Manifold-Manifold Distance(CRMMD) defined as follows:

Given two manifolds $\mathbf{M}^p = \{\mathbf{m}_1^p, \mathbf{m}_2^p, \cdots, \mathbf{m}_m^p\}$ and $\mathbf{M}^q = \{\mathbf{m}_1^q, \mathbf{m}_2^q, \cdots, \mathbf{m}_n^q\}$, the proposed CRMMD between $\mathbf{M}^p$ and $\mathbf{M}^q$ is defined as:

$$d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q) = \min\left\{\min_i d(\mathbf{m}_i^p, \mathbf{M}^q), \min_j d(\mathbf{m}_j^q, \mathbf{M}^p)\right\}, \tag{5.1}$$

where $i \in \{1, 2, \cdots, m\}$, $j \in \{1, 2, \cdots, n\}$ and $d(\mathbf{m}_i^p, \mathbf{M}^q)$ is the distance between local model $\mathbf{m}_i^p$ and manifold $\mathbf{M}^q$, which can be obtained by solving a constraint least square fitting problem:

$$\begin{aligned} d(\mathbf{m}_i^p, \mathbf{M}^q) &= \min_{\mathbf{w}_i} ||\mathbf{m}_i^p - \hat{\mathbf{m}}_i^p||^2 \\ &= \min_{\mathbf{w}_i} ||\mathbf{m}_i^p - \mathbf{M}^q \cdot \mathbf{w}_i||^2, \\ &\quad s.t. \quad \mathbf{1}^T \mathbf{w}_i = 1, \ ||\mathbf{w}_i||_{l^0} = k \end{aligned} \tag{5.2}$$

where $\mathbf{m}_i^p$ denotes the i-th local model from manifold $\mathbf{M}^p$, $\hat{\mathbf{m}}_i^p = \mathbf{M}^q \cdot \mathbf{w}_i$ is an approximation of $\mathbf{m}_i^p$ which is reconstructed by using the local models located on manifold $\mathbf{M}^q$. From the viewpoint of $\mathbf{m}_i^p$, $\mathbf{M}^q$ is a codebook consisting of a set of local models as basis and $\mathbf{w}_i$ is the code for reconstructing $\mathbf{m}_i^p$. We define $\mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbf{R}^n$, thus $\mathbf{1}^T \mathbf{w}_i = 1$ means the summation across all elements in $\mathbf{w}_i$ equals to 1. This balances the importance among different basis that are used for the reconstruction. $||\mathbf{w}_i||_{l^0} = k$ means that there are k non-zero elements in $\mathbf{w}_i$.

Depending on different values of k, we consider three conditions as follows:

(i) When $k = 1$, $\hat{\mathbf{m}}_i^p$ is the nearest one of all the local models that are located on manifold $\mathbf{M}^q$;

(ii) When $1 < k < n$, $\hat{\mathbf{m}}_i^p$ is the collaborative reconstruction from the k nearest neighbors of all the local models that are located on manifold $\mathbf{M}^q$;

(iii) When $k = n$, $\hat{\mathbf{m}}_i^p$ is the collaborative reconstruction from all the local models located on manifold $\mathbf{M}^q$.

In the following, we analyze the properties of the proposed Collaborative Reconstruction-Based Manifold-Manifold Distance (CRMMD) approach under the three different conditions.

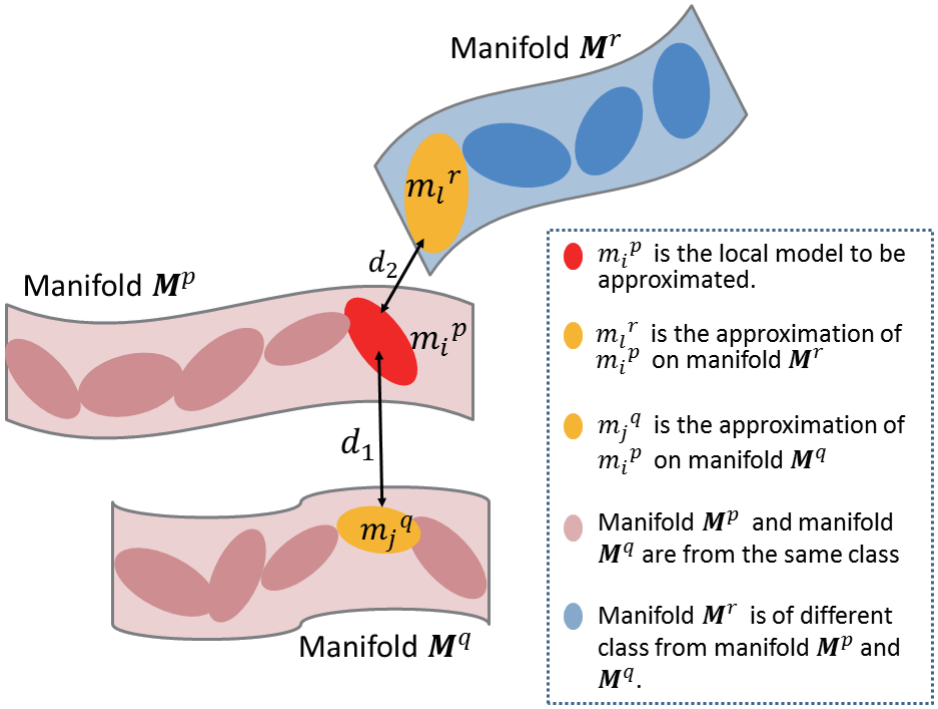

Figure 5.2: Illustration of calculating the CRMMD between manifolds $\{\mathbf{M}^p, \mathbf{M}^q\}$ and the CRMMD between manifolds $\{\mathbf{M}^p, \mathbf{M}^r\}$ when $k = 1$. $\mathbf{m}_j^q$ and $\mathbf{m}_l^r$ are the approximations of $\mathbf{m}_i^p$ on manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^r$ respectively. Thus, $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q) = d(\mathbf{m}_i^p, \mathbf{m}_j^q) = d_1$. Similarly, $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^r) = d(\mathbf{m}_i^p, \mathbf{m}_l^r) = d_2$. However, the CRMMD is sensitive to outliers when $k = 1$, e.g., manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^r$ are from different classes but the inter-class distance d_2 is smaller than the intra-class distance d_1 due to the outlier $\mathbf{m}_l^r$.

5.2.1 Approximation 1 : the Nearest Neighbor

When $k = 1$, the CRMMD becomes:

$$d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q) = \min_{i,j} d(\mathbf{m}_i^p, \mathbf{m}_j^q) = \min_{i,j} ||\mathbf{m}_i^p - \mathbf{m}_j^q||^2, \tag{5.3}$$

In this case, the approximation $\hat{\mathbf{m}}_i^p$ of the local model $\mathbf{m}_i^p$ is represented by using only one single basis in the codebook $\mathbf{M}^q$. The single basis is the nearest one of all local models that are located on manifold $\mathbf{M}^q$. We illustrate the scenario of the CRMMD when $k = 1$ in Fig. 5.2. Under this condition, the CRMMD measures the similarity between two manifolds through computing the distance of their most similar parts, i.e., the minimal distance of all the pairwise distances among local models from the two manifolds is

selected as the CRMMD. It is observed that the manifold-to-manifold distance utilized in the popular method MMD [16] can be considered as the special case of our proposed CRMMD method when $k = 1$.

However, the CRMMD may be sensitive to outliers when $k = 1$. We illustrate this in Fig. 5.2. On the manifold $\mathbf{M}^p$, the local model $\mathbf{m}_j^q$ is the nearest neighbor of the local model $\mathbf{m}_i^p$. On the manifold $\mathbf{M}^r$, the local model $\mathbf{m}_l^r$ is the nearest neighbor of the local model $\mathbf{m}_i^p$. Thus we use the $\mathbf{m}_j^q$ and $\mathbf{m}_l^r$ as the approximations of the local model $\mathbf{m}_i^p$. Then the CRMMD between manifolds $\mathbf{M}^p$ and $\mathbf{M}^q$ becomes:

$$d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q) = d(\mathbf{m}_i^p, \mathbf{m}_j^q) = ||\mathbf{m}_i^p - \mathbf{m}_j^q||^2 = d_1, \tag{5.4}$$

and the CRMMD between manifolds $\mathbf{M}^p$ and $\mathbf{M}^r$ becomes:

$$d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^r) = d(\mathbf{m}_i^p, \mathbf{m}_l^r) = ||\mathbf{m}_i^p - \mathbf{m}_r^l||^2 = d_2, \tag{5.5}$$

Since the manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^q$ are from the same class and the manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^r$ are from different classes, the intra-class distance $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q)$ should be smaller than the inter-class distance $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^r)$. However, as illustrated in Fig. 5.2, the $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^q)$ is greater than the $d_{CRMMD}(\mathbf{M}^p, \mathbf{M}^r)$. This may be caused by the outlier $\mathbf{m}_r^l$ that reduces the inter-class distance between the manifolds $\mathbf{M}^p$ and $\mathbf{M}^r$.

5.2.2 Approximation 2 : Reconstruction from the k Nearest Neighbors

When $1 < k < n$, the approximation $\hat{\mathbf{m}}_i^p$ of the local model $\mathbf{m}_i^p$ is collaboratively reconstructed from the k nearest neighbors of all the local models that are located on manifold $\mathbf{M}^q$. Thus, we solving the constrained least square fitting problem involved in the $d(\mathbf{m}_i^p, \mathbf{M}^q)$ of the CRMMD as follows:

$$\begin{aligned} d(\mathbf{m}_i^p, \mathbf{M}^q) &= \min_{\mathbf{w}_i} ||\mathbf{m}_i^p - \hat{\mathbf{m}}_i^p||^2 \\ &= \min_{\mathbf{w}_i} ||\mathbf{m}_i^p - \sum_{l=1}^{k} \mathbf{w}_{i,l} \mathbf{m}_l^q||^2, \\ &\quad s.t. \quad \mathbf{1}^T \mathbf{w}_i = 1. \end{aligned} \tag{5.6}$$

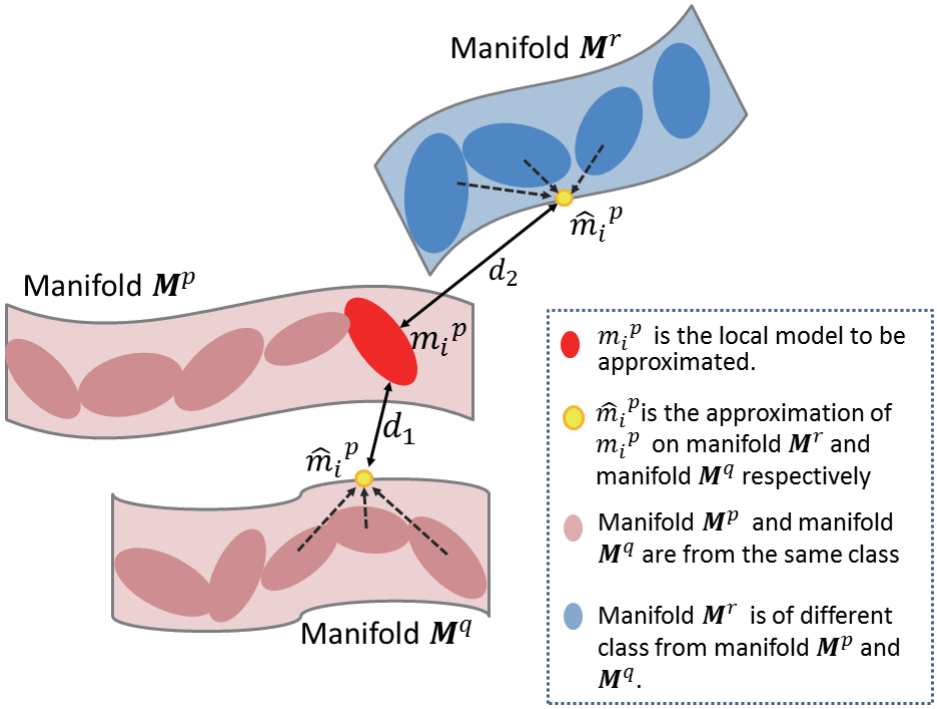

Figure 5.3: Illustration of calculating the CRMMD between manifolds $\{\mathbf{M}^p, \mathbf{M}^q\}$ and the CRMMD between manifolds $\{\mathbf{M}^p, \mathbf{M}^r\}$ when $1 < k < n$. The approximations of $\mathbf{m}_i^p$ on manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^r$ are linear combinations of multiple bases. Thus, in the CRMMD, $d_1, d_2 = d(\mathbf{m}_i^p, \hat{\mathbf{m}}_i^p)$. In this case, the CRMMD becomes more robust to outliers such that the intra-class distance d_1 is smaller than the inter-class distance d_2.

In this case, each local model $\mathbf{m}_i^p$ is approximated by the linear combination of the k nearest basis in the codebook $\mathbf{M}^q$. The correlation information across the multiple basis can be used to provide a more robust approximation $\hat{\mathbf{m}}_i^p$, since the local models with good qualities are able to alleviate the perturbations introduced by noisy local models. As shown in Fig. 5.3, the approximations of $\mathbf{m}_i^p$ on manifold $\mathbf{M}^p$ and manifold $\mathbf{M}^r$ are linear combinations of multiple basis in each codebook. It can be observed that the proposed CRMMD becomes more robust to outliers such that the intra-class distance d_1 is smaller than inter-class distance d_2.

Under this condition, the CRMMD has an analytical solution which can be derived as follows:

$$\mathbf{w}_i = \left[\left(\tilde{\mathbf{M}}_s^q\right)^T \tilde{\mathbf{M}}_s^q + \mu\mathbf{I}\right] \setminus \mathbf{1}, \tag{5.7}$$

$$\mathbf{w}_i^* = \mathbf{w}_i/(\mathbf{1}^T\mathbf{w}_i), \tag{5.8}$$

where μ is the regularization parameter, $\mathbf{I}$ is the identity matrix, $\tilde{\mathbf{M}}_s^q = \mathbf{M}_s^q - \mathbf{m}_i^p \cdot \mathbf{1}^T$, and $\mathbf{M}_s^q$ contains the k nearest neighbors of $\mathbf{m}_i^p$ located on $\mathbf{M}^q$.

5.2.3 Approximation 3 : Reconstruction from All the Local Models

When $k = n$, the approximation $\hat{\mathbf{m}}_i^p$ of the local model $\mathbf{m}_i^p$ is collaboratively reconstructed from all the local models $\mathbf{m}_j^q$ that are located on manifold $\mathbf{M}^q$. Thus, the $d(\mathbf{m}_i^p, \mathbf{M}^q)$ of the CRMMD can be obtained through solving the least square fitting problem with less constraint as follows:

$$\begin{aligned} d(\mathbf{m}_i^p, \mathbf{M}^q) &= \min_{\mathbf{w}_i} ||\mathbf{m}_i^p - \mathbf{M}^q \cdot \mathbf{w}_i||^2, \\ &\quad s.t. \quad \mathbf{1}^T\mathbf{w}_i = 1. \end{aligned} \tag{5.9}$$

In this case, each local model $\mathbf{m}_i^p$ is approximated by collaboratively combining all the basis in the codebook $\mathbf{M}^q$. Similarly, the analytical solution of the CRMMD when $k = n$ can be obtained as below:

$$\mathbf{w}_i = \left[\left(\tilde{\mathbf{M}}^q\right)^T \tilde{\mathbf{M}}^q + \mu\mathbf{I}\right] \setminus \mathbf{1}, \tag{5.10}$$

$$\mathbf{w}_i^* = \mathbf{w}_i/(\mathbf{1}^T\mathbf{w}_i), \tag{5.11}$$

where μ is the regularization parameter and $\tilde{\mathbf{M}}^q = \mathbf{M}^q - \mathbf{m}_i^p \cdot \mathbf{1}^T$.

Under this condition, the CRMMD becomes not only a robust distance measure since all the basis are utilized in the collaborative reconstruction which provides correlation information across multiple basis, but also an efficient distance measure because there is no pre-procedure for searching the k nearest neighbors of $\mathbf{m}_i^p$ in the codebook $\mathbf{M}^q$.

5.3 Experiments

In the experimental section, we evaluate the proposed Collaborative Reconstruction-Based Manifold-to-Manifold Distance (CRMMD) method on the task of face recognition based on image sets under three conditions: $k = 1$, $1 < k < n$ and $k = n$.

5.3.1 Experimental Setup

We conduct the experiments under several configurations described as follows:

1. Types of local models: We use 2 different approaches to describe a local model in our experiments.

- **MLP**: Similar to the local models introduced in Section 4.4.1, the Maximal Linear Patch (MLP) proposed in the reference [16] is a linear subspace that is spanned by data points in the subset. Thus, each nonlinear manifold is represented by using a collection of MLPs.

- **k-means clustering**: k-means clustering [124] is also utilized in our experiments, since it is able to provide stable and representative mean vectors for describing local models. We set the value of k similar to the experimental settings in the previous chapter: With the number of face images in a set changing from 8 to 1000, the value of k varies in the range of $[3, 29]$ in our experiments.

2. Types of manifold-to-manifold distance: To illustrate the efficacy of our proposed CRMMD method, we compare it with the existing manifold-to-manifold distance, i.e., NN-MMD.

- **NN-MMD**: When each manifold has been represented by using multiple local models, the pairwise distances between local models from two different manifolds are computed and the distance with the smallest value is selected and considered as the dissimilarity of the two manifolds. For easy of explanation, we denote this manifold-to-manifold distance as the Nearest Neighbor-based MMD (NN-MMD).

- **Proposed CRMMD**: We test the performance of the proposed CRMMD under three conditions: $k = 1$, $1 < k < n$ and $k = n$. In the experiments, we divided each image set into around 20 subsets (local models). When $1 < k < n$, we adopt $k = n/2$ which is selected through a series of experiments with varying k $(k = 1, n/10, 2n/10, \cdots, 9n/10, n)$.

3. Other unsupervised set-based matching methods: In this chapter, all the methods we have investigated are unsupervised approaches that only consist of two stages: the modeling stage and the matching stage. Thus, in the experimental section, we compare the proposed the CRMMD method with some state of the arts that adopt the unsupervised matching stage based on image sets.

- **The AHISD method**: The Affine hull-based Image Set Distance (AHISD) was proposed in reference [98]. In their approach, they model each image set by using the face images in each set to span a convex hull or an affine hull. Based on these representations, the distance between two image sets can be obtained by computing the distance between the closest points from the two convex geometric regions. The source code of the AHISD approaches is provided by original authors. In the implementations of the AHISD method, there is no parameter to be tuned.

- **The MMD method**: In the previous chapter, the framework of the Manifold-to-Manifold Distance method (MMD) [16] has been introduced in Section 4.2. As one of the unsupervised set-based matching methods, MMD is utilized in our experiments to compare with the proposed CRMMD method. It is noted that the MMD is a special case of the proposed CRMMD, i.e., MMD equals to CRMMD when $k = 1$. The source code of the MMD approach is provided by original authors and we follow the parameter settings as those in their papers to conduct our experiments.

4. The three databases: In the following experiments, we will conduct all the approaches as mentioned above on three widely used video face databases, where each video sequence is saved as an image set.

- **The Honda/UCSD Database**: As introduced in Section 4.4.1, there are 59 video sequences from 20 different individuals in the Honda/UCSD Database [127], where each individual has at least two video sequences. The qualities of videos is the best among the three databases due to it is obtained under controlled conditions. There are some exemplar cropped face images from the Honda/UCSD database are shown in Fig. 4.3.

- **The CMU MoBo Database**: As introduced in Chapter 3, there are 96 videos sequences from 24 different persons in the CMU MoBo database [119]. Each person has 4 video sequences containing a large number of variations in poses and expressions. Some exemplar face images from the CMU MoBo database are shown in Fig. 3.9.

- **The Youtube Celebrities Database**: As introduced in Section 4.4.1, the Youtube Celebrities Database [128] contains 1910 video sequences from 47 different individuals. The videos in Youtube Celebrities database are with low qualities due to obtained from the real-life scenarios. Some well-cropped face images from the Youtube Celebrities database are shown in Fig. 4.4.

5.3.2 Experimental Results and Analysis

We conduct 3 sets of experiments as follows: 1. Using the MLP to construct each subset for all the methods; 2. Using the k-means clustering to construct each subset for all the methods; 3. Studying the effect of k.

5.3.2.1 Experimental Results of Methods Using MLP

The experimental results on the Honda/UCSD Database: To conduct fair comparisons, we use the MLP method to construct each subset for the two different manifold-to-manifold distance measures: the NN-MMD and the proposed CRMMD. On the Honda/UCSD database, We apply the Viola-Jones face detection algorithm [47] to detect face regions. Then each face image is cropped to 20×20 pixels. We adopt the same settings as described in Section 4.4.2 where 50 frames, 100 frames and full length of each video

sequence are used in the experiments respectively. Table 5.1 illustrates the experimental results of the two different distance measures based on MLP on the Honda/UCSD database.

Table 5.1: Face recognition results of two manifold-to-manifold distance measures using MLP on the Honda/UCSD database.

Methods	50 frames	100 frames	Full length	Average
MLP+NN-MMD [16]	82.05%	82.05%	94.87%	86.32%
MLP+CRMMD	82.05%	84.62%	94.87%	87.20%

Without analyzing the experimental results, we compare the above two manifold-to-manifold distance measures on the CMU MoBo database and Youtube Celebrities Database as follows. After that, we will discuss about the observations obtained from these experimental results together.

The experimental results on the CMU MoBo Database: Similar to the previous experiment, we compare the NN-MMD method and the proposed CRMMD method on the CMU MoBo database. In this database, each person has 4 video sequences and the Viola-Jones face detection algorithm [47] is applied to detect and crop faces in each frame. Each time we randomly select one image set per person to used as gallery sets and the rest three image sets per person are used as probe sets. The experiments are repeated 30 times and the averaged results are summarized in Table 5.2.

Table 5.2: Face recognition results of two manifold-to-manifold distance measures using MLP on the CMU MoBo database.

Methods	Recognition Results
MLP+NN-MMD [16]	$93.05\% \pm 1.48\%$
MLP+CRMMD	$94.44\% \pm 1.81\%$

The experimental results on the Youtube Celebrity Database: We first apply the incremental visual tracking algorithm [129] to track face regions across frames of each video by using the initials that are manually labeled. We crop the tracked face region of each frames and resize it to 50×50 pixels. Then we conduct five-fold cross validation experiments that described in Section 4.4.3. We repeat the random division of five folds 50 times and record the averaged recognition rates and standard deviations in Table 5.3.

Table 5.3: Face recognition results of two manifold-to-manifold distance measures using MLP on the Youtube Celebrity database.

Methods	Recognition Results
MLP+NN-MMD [16]	$54.78\% \pm 2.16\%$
MLP+CRMMD	$56.03\% \pm 1.59\%$

Analysis of the experimental results on the three databases: From the experimental results on the above three databases, it is observed that the recognition performance of the proposed CRMMD only slightly outperforms the NN-MMD method. The phenomenons are consistent across different databases where the MLP is used to construct subsets. The reason may be that the MLP is not an appropriate way to cluster the subsets for CRMMD. The MLP aims to find a subset that is able to strictly span a linear subspace. However, we describe this subset by using its mean vector, not a linear subspace. This leads to that the advantage of the MLP is not utilized in our CRMMD method. Due to the high computational complexities of linear subspace construction and matching, we use the mean vector as the local model instead of using a linear subspace. Thus we need adopt a clustering method that is able to divide an image set into several subsets, each of which is suitable for computing a stable and representative mean vector. Based on these analysis, we adopt the k-means clustering method to divide each image set in the following experiments.

5.3.2.2 Experimental Results of Methods Using k-means

The experimental results on the Honda/UCSD Database: In this experiment, we adopt the configuration described in [117] [98] [127] where there are 20 video sequences are selected to be stored in the gallery dataset and the remaining 39 videos are used for testing. We evaluate the AHISD, MMD and the proposed CRMMD method under three settings: randomly selecting 50 frames from each video, randomly selecting 100 frames from each video and using all the frames of each video. Under each setting, we repeat the random selection 50 times and average all the results as the final recognition rate. The performances of the three methods are shown in Table 5.4.

In the Table 5.4, the average performance of the proposed CRMMD ($1 < k < n$) achieves 89.20% which outperforms those of MMD and AHISD methods. Comparing with

Table 5.4: Experimental results of the three comparative methods using k-means clustering on the Honda/UCSD database.

Methods	Average
MMD	$86.53\% \pm 7.25\%$
AHISD	$87.17\% \pm 2.60\%$
CRMMD $(k = 1)$	$85.23\% \pm 7.78\%$
CRMMD $(1 < k < n)$	$\mathbf{89.20}\% \pm 5.75\%$
CRMMD $(k = n)$	$88.46\% \pm 5.46\%$

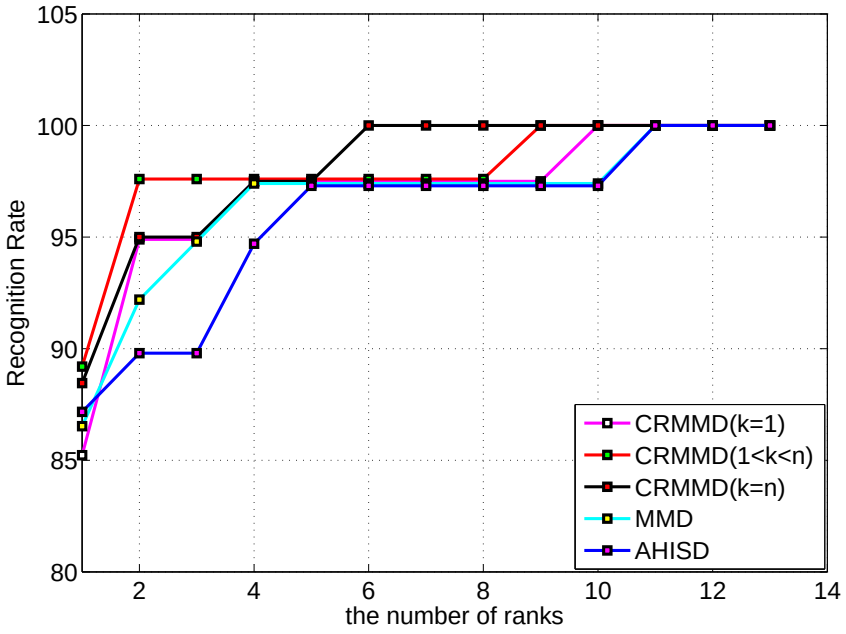

Figure 5.4: The rank r recognition rates of the five comparative approaches on the Honda/UCSD database.

AHISD method, CRMMD models each manifold by using multiple local models which describes an image set better than a single model. Comparing with MMD, CRMMD $(1 < k < n$ and $k = n)$ adopts multiple bases to reconstruct a local model instead of using a single basis as approximation. The performance of CRMMD $(k = 1)$ shows the lowest recognition rate, since it is sensitive to outliers when CRMMD uses the nearest neighbor as the approximation.

We also illustrate the rank r recognition rates of the five comparative approaches in Fig. 5.4 where the rank $r \in [1, 14]$. We can observe from this figure that the proposed

Table 5.5: Experimental results of the three comparative methods using k-means clustering on the CMU MoBo database.

Methods	Average
MMD	$94.40\% \pm 3.00\%$
AHISD	$94.39\% \pm 1.90\%$
CRMMD ($k = 1$)	$97.10\% \pm 4.04\%$
CRMMD ($1 < k < n$)	$97.78\% \pm 3.94\%$
CRMMD ($k = n$)	$\mathbf{98.52}\% \pm 3.05\%$

CRMMD methods consistently outperform the other methods when the rank $r \geq 2$.

The experimental results on the CMU MoBo Database: On the CMU MoBo Database, we divide each face image into multiple patches (size of 8×8 pixels) and then extract the uniform LBP feature from each patch. A face image is represented by using a feature vector that is concatenated by the LBP descriptors extracted from each local patch. Then we adopt the experimental configuration as follows: for each person, one image set is randomly selected to be stored in the gallery set and the other three image sets are used as probe sets. We repeat the random divisions 30 times for each method. The average results are recorded in Table 5.5.

From the results shown in Table 5.5, the average performance of our proposed CRMMD method ($k = n$) achieves the best recognition rate 98.52% which outperforms MMD by 4.12% and AHISD by 4.13%. Comparing with the Honda/UCSD database, videos in the CMU MoBo database contain more noises and larger variations in pose which leads to outliers in manifolds. Under this situation, the proposed CRMMD shows its efficacy and robustness to outliers.

As investigated in Chapter 3, the best recognition results of the single model-based methods, including the MSM method, the SD method, the WSD method, the proposed FOWSD and EWSD methods, is stable around 93% on the CMU MoBo database. In this chapter where we adopt the multi-model based framework, it is noted that the best recognition performance of the multi-model based methods (the MMD method and the proposed CRMMD method) achieves 98.52% on the CMU MoBo database, which is much better the performance of the single model-based methods. Thus, it is better to describe an image set by using multiple local models when this image set is with nonlinear distribution.

Table 5.6: Experimental results of the three comparative methods using k-means clustering on the YouTube Celebrities database.

Methods	Average
MMD	$54.78\% \pm 4.66\%$
AHISD	$57.13\% \pm 5.43\%$
CRMMD ($k = 1$)	$55.52\% \pm 5.63\%$
CRMMD ($1 < k < n$)	$\mathbf{57.40}\% \pm 5.62\%$
CRMMD ($k = n$)	$57.22\% \pm 6.02\%$

The experimental results on the YouTube Celebrities Database: By applying a tracking algorithm [129], we successfully tracked over 80% faces of the videos and resized the face to 50×50 for each frame. We follow the experimental configuration mentioned in [117] to conduct five-fold cross validation experiments. The 1910 video sequences are divided into five folds, each of which contains 423 videos from 47 persons with 9 videos per person. In each fold, 3 image sets per person are randomly selected to be stored in the gallery set, and the remaining 6 videos per person are used as probe sets. We repeat the random division 50 times and record the averaged recognition rates and standard deviations in Table 5.6.

In Table 5.6, it is observed that the performance of the proposed CRMMD method ($1 < k < n$) reaches the best recognition rate 57.40%, which is better than the MMD by 2.62% and slightly higher than AHISD 57.13%. Compared with the MMD method, CRMMD method adopts collaborative reconstruction to handle outliers, which leads to the better performance of the CRMMD method than that of the MMD method. Compared with the AHISD method, the experimental results shows that the performances of CRMMD and AHISD are comparable on this challenging database. In the YouTube Celebrities database, the number of frames of each video sequence varies in a large range. When the size of a subset is not very large, the mean vector is a good choice to represent this subset. However, when the size of a subset in too large, the mean vector may not be a sufficient descriptor for a large or sparse subset. In this situation, the AHISD method that utilizes an affine hull or a convex hull to characteristic an image set is able to achieve an acceptable recognition performance.

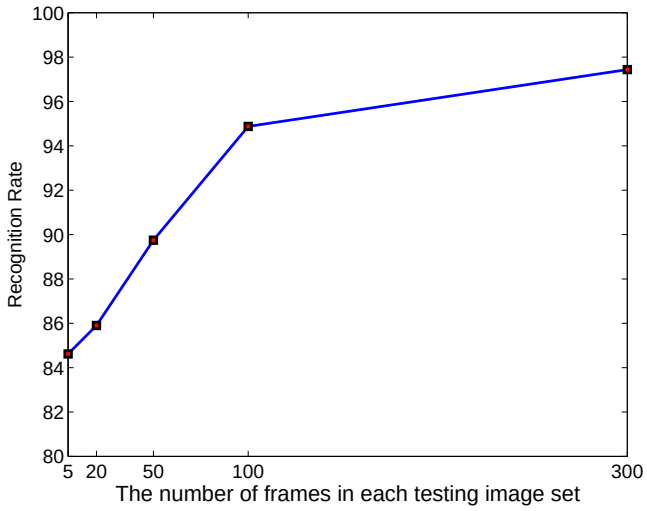

Figure 5.5: The recognition rate goes up with the increasing of the number of face images in each probe set.

5.3.2.3 The effect of the Number of Images in a Probe Set

To demonstrate that the efficacy of the methods for the task of face recognition based on image sets, we design an experiment on the Honda/UCSD database as follows: We use all the images of each gallery video to construct the gallery set and use the first f frames of each probe video to store as each probe set where $f \in [5, 300]$. Under this configuration, we increase the number of images for each probe set and keep the number of images per gallery set unchanged. The CRMMD method is used in this experiment and the recognition results are illustrated in Fig.5.5.

From the Fig.5.5, it can be observed that: When the gallery sets contain sufficient face images for each person, the recognition performance of the FRBIS methods goes up with the increasing of the number of face images in each probe set. The recognition rate is only 84.62% when the number of images per set is 5. The recognition rate rises up to 94.87% when the number of images per set becomes 50. The recognition rate achieves 97.44% when the number of images per set is greater than 100. This experiment verifies the efficacy of methods that recognize a person through a set of face images rather than using a single image.

5.3.2.4 The effect of k

In the proposed CRMMD method, the parameter k denotes the number of local models used in collaborative reconstruction. It is important for us to clarify that how the various values of k can affect the recognition performance of the proposed CRMMD method. We conduct a series of experiments with varying $k(k = 1, n/10, 2n/10, \cdots, 9n/10, n)$ based on the Honda/UCSD database where n denotes the number of basis in the codebook. We perform the CRMMD method under three settings as mentioned in previous experiments. Using the first 50 frames, the first 100 frames and full length of each video sequence. Under each setting, we repeat the experiment 30 times and average the recognition rates over the three setting conditions. Table 5.7 summarized the averaged results as follows.

Table 5.7: Face recognition results of the CRMMD method with various values of k on the Honda/UCSD database.

Values of k	$k = 1$	$k = n/10$	$k = 2n/10$	$k = 3n/10$
Recognition Rates	85.47%	85.61%	85.04%	85.89%
Values of k	$k = 4n/10$	$k = 5n/10$	$k = 6n/10$	$k = 7n/10$
Recognition Rates	88.03%	**89.32**%	88.20%	88.46%
Values of k	$k = 8n/10$	$k = 9n/10$	$k = n$	
Recognition Rates	88.20%	89.07%	85.04%	

From the Table 5.7, it is observed that the recognition results of CRMMD are not sensitive to k when it takes values in the range of $[4n/10, 9n/10]$ and the best recognition rate of 89.32% is achieved at $k = n/2$. Thus, we chose $k = n/2$ in our experiments.

5.4 Conclusion

In this chapter, we have investigated the stage of manifold-to-manifold distance measure under the framework of multi-model based methods, as a continuous of the stage of local model construction discussed in the previous chapter. We first introduced the limitations of the existing manifold-to-manifold distance measure which we denote as NN-MMD in Section 5.1. In view of these limitations, we have proposed a more robust and flexible manifold-to-manifold distance measure called collaborative reconstruction-based manifold-manifold distance (CRMMD) for face recognition based on image sets

in Section 5.2. The CRMMD between two manifolds can be obtained by computing the distance between each local model and its approximation which was reconstructed from neighbors on the other manifold. This reconstruction increased robustness of the CRMMD method. In Section 5.3, we verified the efficacy of the proposed CRMMD method through comparing with the state of the arts that are based on the unsupervised set-based matching on three widely used databases.

It is noted that the methods investigated in Chapter 4 and Chapter 5 are based on unsupervised set-based matching. Different from this, we will continue to discuss the multi-model based methods in a supervised manner in the next chapter.

Chapter 6

Multi-Manifold Metric Learning Approach

6.1 Introduction

In Chapter 4 and Chapter 5, it is observed that the multi-model based methods outperform the single model-based methods from the experimental results. Thus, we continue to improve the approaches based on the multiple models in Chapter 6. The motivations of our work in this chapter are explained as below:

- **From the viewpoint of supervised learning**: It is known that the supervised learning is very useful for the pattern classification task, since the discriminative information provided by class labels are fully utilized in the training phase. Thus, we add a discriminative learning step before the stage of manifold-to-manifold distance measure to further enhance the performance of the multi-model based methods in this work. Furthermore, the existing discriminative learning algorithms usually learn only one low-dimensional subspace for all the samples from different classes. However, the distributions of manifolds in different classes are not consistent, which leads to unappropriate mappings. In view of this, we aim to design a novel learning approach that seeks a collection of discriminative mappings, one for each class, to adaptively transform the manifolds in different classes.

- **From the viewpoint of metric learning**: Recently, a number of distance metric learning methods have been proposed in references [131] [132] [133] [134] [135] [136].

Most of them learn a distance metric from single images and cannot deal with image sets. A possible solution is to make use of each image sample within a set separately and then learn a distance metric from these image samples. However, the relations of face samples within a set is missed and some discriminative information is ignored. In this work, we model each set as a manifold and then learn distance metrics from these manifolds rather than single images, which is more suitable for our face recognition from image sets problem.

In this chapter, we propose a multi-manifold metric learning (MMML) method for the task of face recognition based on image sets. Different from most existing metric learning algorithms that learn the distance metric based on single-shot images, our method aims to learn distance metrics to measure the similarity of pairs of manifolds. Given an image set, we first model it as a manifold and learn multiple distance metrics under which the intra-class manifold variations are minimized and inter-class manifold variations are maximized, simultaneously. For each person, we learn a distance metric by using such criterion so that the learned distance metrics are person-specific and more discriminative. In the classification phase, a probe image set is also modeled as a collection of affine hulls and compared with each gallery set associated with the learned distance metric. Fig. 6.1 shows the working flow of our approach. There are three nice properties of the proposed MMML algorithm summarized as follows:

- MMML learns a discriminant metric which aims to maximize the separability of neighbor manifolds from different classes, at the same time the compactness of manifolds from the same class is minimized.

- MMML learns a class-specific metric. In contrast to existing learning algorithms which learn only one distance metric or a global linear transformation for all the samples from different classes, the proposed method learns a collection of discriminative distance metrics, one for each class. This class-specified distance metric is able to better differentiate the current class from others than a single metric, which may be not discriminative enough because face images from different classes may lie on different manifolds with different intrinsic dimensions.

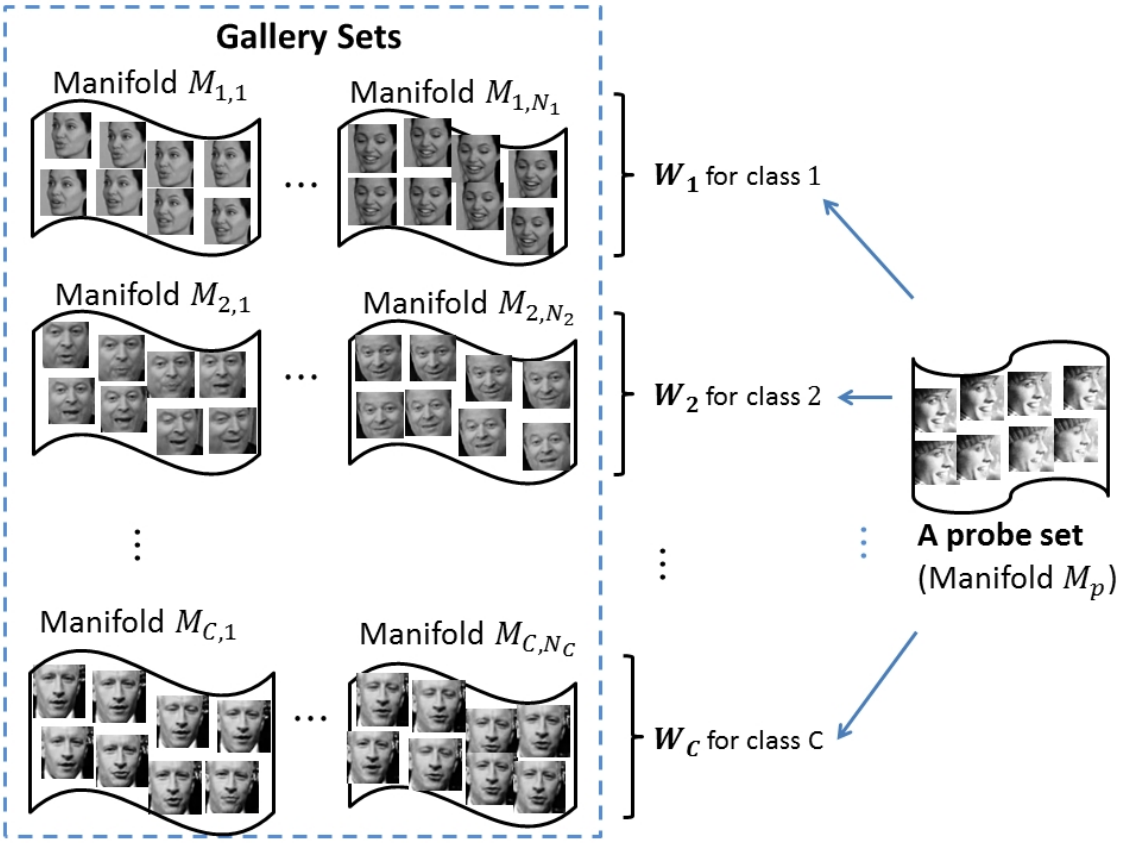

Figure 6.1: Illustration of our proposed multi-manifold metric learning method. For each image set, we model it as a manifold, which is further approximated as a collection of affine hulls. For each person, we learn a person-specific distance metric to maximize the inter-class manifold variations and minimize intra-class manifold variations, simultaneously, such that more discriminative information can be exploited for recognition. In the test phase, the test image set is compared with each gallery image set associated with the learned distance metric, and a label is assigned according to the nearest neighbor rule.

- MMML learns a set-based metric. The previous metric learning algorithms are instance-based metric, which cannot deal with the comparisons among image sets. Thus, in the scenario of classification based on image sets, the MMML distance metric is more suitable to measure the distances between manifolds.

The rest of this chapter is organized as follows: Based on the analysis mentioned above, we propose a Multi-Manifold Metric Learning (MMML) method in Section 6.2, which utilizes the discriminative property of a metric learning step to enhance the per-

formance of multi-model based methods. Our method is extensively evaluated on three popular face databases and compared to the state of the arts. Experimental results are presented to show the effectiveness of the proposed method in Section 6.3, and Section 6.4 concludes the investigations of this chapter.

6.2 Proposed Approach

6.2.1 Multi-Affine Hulls Model

Consider N training sets from C classes. For each set, the face images are assumed to lie on or near an underlying nonlinear and intrinsically low-dimensional manifold $\mathbf{M}$. To better capture the characteristic of each manifold, we adopt multiple local models rather than a single model. Existing multi-model based methods [16, 118] first divide a nonlinear manifold $\mathbf{M}$ into several subsets $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_m\}$ and each subset $\mathbf{s}_i$, $i \in \{1, 2, \cdots, m\}$, is then described by a linear subspace, such that a collection of linear subspaces is used to model the manifold $\mathbf{M}$. However, considering that the linear subspace is a rather loose approximation of a subset, we use more compact representations, i.e., a collection of affine hulls to characterize a manifold $\mathbf{M} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_m\}$, where $\mathbf{h}_i$, $i \in \{1, 2, \cdots, m\}$, is an affine hull. Since the mean information is important to characterize an affine hull $\mathbf{h}_i$ because it reflects the averaged positions of face images within a set, we divide an image set into multiple subsets by using the spectral clustering algorithms [125] [126] such that a stable and representative mean vector can be obtained for each affine hull.

The spectral clustering: The spectral clustering algorithms [125] [126] cluster the data points based on the first k eigenvectors of a normalized Laplacian matrix $\mathbf{L}$ of $\mathbf{S}$, which is an adjacency matrix storing the similarities between pairwise points of a data set $\mathbf{X}$. Then the objective function aims to find a collection of orthogonal embeddings of the data set denoted as $\mathbf{V}$, which consists of the approximations of the cluster indicator vectors:

$$\max_{\mathbf{V} \in \mathbf{R}^{n \times k}} tr\left(\mathbf{V}^T \mathbf{L} \mathbf{V}\right) \qquad s.t. \;\; \mathbf{V}^T \mathbf{V} = \mathbf{I}, \tag{6.1}$$

where the optimal $\mathbf{V}$ contains the first k eigenvectors of the symmetric normalized Laplacian matrix $\mathbf{L}$ as columns. Then each row vector of $\mathbf{V}$ can be normalized and used to assign a data point to one of the k clusters through the k-means algorithm [124]. The

spectral clustering algorithm has many nice properties such as well-defined mathematic framework, more discriminative nature for clustering, good performance on clusters with arbitrary shapes, etc.

The affine hull model for each subset: For a subset $\mathbf{s}_i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \cdots, \mathbf{x}_{ni}^i]$, its affine hull approximation $\mathbf{h}_i$ is defined as

$$\mathbf{h}_i = \left\{ \mathbf{x}_i = \mu_i + \mathbf{U}_i \mathbf{v}_i | \mathbf{v}_i \in \mathbf{R}^D \right\}, \tag{6.2}$$

where sample $\mathbf{x}_i \in \mathbf{R}^D$, μ_i is the mean vector of the subset $\mathbf{s}_i$, $\mathbf{U}_i$ is the orthonormal basis that spans the whole affine hull and $\mathbf{v}_i$ is a vector of free parameters of $\mathbf{x}_i$ within the affine hull. $\mathbf{U}_i$ is obtained from the singular vectors of Singular Value Decomposition (SVD) on $[\mathbf{x}_1^i - \mu_i, \mathbf{x}_2^i - \mu_i, \cdots, \mathbf{x}_{ni}^i - \mu_i]$. By using an affine hull, a subset can be compactly represented since the affine subspace is able to reduce more redundant dimensions than a linear subspace that is forced to go through the origin.

Differences from the AHISD method [98]: The method that utilizes affine hulls or convex hulls to represent image sets was first proposed in [98]. In this AHISD method, an image set is modeled by using a single affine hull or convex hull. In contrast, we first cluster an image set into multiple subsets through a spectral clustering method, and model each subset as an affine hull. Then the whole image set is described by using a collection of affine hulls, i.e., $\mathbf{M} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_m\}$. Thus, our proposed method has a stronger capability to characterize an image set with nonlinear distributions.

6.2.2 Affine hull-based Distance Metric

Most existing work such as [104], [16] design the manifold-to-manifold distance measure as the dissimilarity between the most similar parts of two image sets, i.e.,

$$d^{\mathrm{mf}}(\mathbf{M}_p, \mathbf{M}_q) = \min_{\mathbf{h}_i \in \mathbf{M}_p} \min_{\mathbf{h}_j \in \mathbf{M}_q} d^{\mathrm{aff}}(\mathbf{h}_i, \mathbf{h}_j), \tag{6.3}$$

where $d^{\mathrm{mf}}(\mathbf{M}_p, \mathbf{M}_q)$ denotes the distance between manifolds $\mathbf{M}_p$ and $\mathbf{M}_q$, and $d^{\mathrm{aff}}(\mathbf{h}_i, \mathbf{h}_j)$ denotes the distance between affine hulls $\mathbf{h}_i$ and $\mathbf{h}_j$.

Instead of directly comparing the two affine hulls, we compare them by using a learned distance metric. The conventional metric learning algorithms learn a Mahanalobis distance metric that is an instance-based metric. However, in the considered scenario, an

affine hull-based distance metric is needed. Hence, we define a distance metric over affine hulls as:

$$d^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{A}) = \alpha \frac{(\mu_i^T \mathbf{A} \mu_j)^2}{(\mu_i^T \mathbf{A} \mu_i)(\mu_j^T \mathbf{A} \mu_j)} + (1-\alpha) tr(\Lambda), \tag{6.4}$$

where α is a scalar parameter and Λ is obtained from the SVD of $(\tilde{\mathbf{U}}_i^T \mathbf{A} \tilde{\mathbf{U}}_j)$, i.e., $(\tilde{\mathbf{U}}_i^T \mathbf{A} \tilde{\mathbf{U}}_j) = \mathbf{Q}_{ij} \Lambda \mathbf{Q}_{ji}^T$. $\mathbf{U}_i$ and $\mathbf{U}_j$ are orthonormal bases that span two affine hulls respectively. Before we explain $\tilde{\mathbf{U}}_i$ and $\tilde{\mathbf{U}}_j$ that are transformed from $\mathbf{U}_i$ and $\mathbf{U}_j$, we first introduce a decomposition of $\mathbf{A}$, where $\mathbf{A} = \mathbf{W}\mathbf{W}^T$ and $\mathbf{W} \in \mathbf{R}^{D \times d}$. This makes the parameter matrix $\mathbf{A}$ be a $D \times D$ symmetric and positive semidefinite matrix such that the affine hull-based metric is a valid metric that satisfies non-negativity, symmetry and triangle inequality. Then we use $\mathbf{W}\mathbf{W}^T$ instead of $\mathbf{A}$ such that Eq. (6.4) can be rewritten as

$$\begin{aligned}
& d^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{W}) \\
= \; & \alpha \frac{(\mu_i^T \mathbf{W}\mathbf{W}^T \mu_j)^2}{(\mu_i^T \mathbf{W}\mathbf{W}^T \mu_i)(\mu_j^T \mathbf{W}\mathbf{W}^T \mu_j)} + (1-\alpha) tr(\Lambda) \\
= \; & \alpha \frac{\left[\left(\mathbf{W}^T \mu_i\right)^T \left(\mathbf{W}^T \mu_j\right) \right]^2}{\left[\left(\mathbf{W}^T \mu_i\right)^T \left(\mathbf{W}^T \mu_i\right) \right] \left[\left(\mathbf{W}^T \mu_j\right)^T \left(\mathbf{W}^T \mu_j\right) \right]} \\
+ \; & (1-\alpha) tr(\Lambda)
\end{aligned} \tag{6.5}$$

where Λ is obtained from the SVD: $(\tilde{\mathbf{U}}_i^T (\mathbf{W}\mathbf{W}^T) \tilde{\mathbf{U}}_j) = (\mathbf{W}^T \tilde{\mathbf{U}}_i)^T (\mathbf{W}^T \tilde{\mathbf{U}}_j) = \mathbf{Q}_{ij} \Lambda \mathbf{Q}_{ji}^T$. It is observed that Λ contains the principal angles between matrices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$ when the two matrices are orthonormal basis matrices of subspaces. Although $\mathbf{U}_i$ and $\mathbf{U}_j$ are orthonormal matrices, in the subspace $\mathbf{W}$, $(\mathbf{W}^T \mathbf{U}_i)$ and $(\mathbf{W}^T \mathbf{U}_j)$ are not generally orthonormal basis matrices. Thus, to obtain the orthogonalized matrices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$, we orthogonalize $(\mathbf{W}^T \mathbf{U}_i)$ and $(\mathbf{W}^T \mathbf{U}_j)$ by applying the QR decomposition

$$\begin{aligned}
\mathbf{W}^T \mathbf{U}_i &= \Phi_i \mathbf{R}_i \\
\mathbf{W}^T \mathbf{U}_j &= \Phi_j \mathbf{R}_j
\end{aligned} \tag{6.6}$$

where Φ_i and Φ_j are orthogonal matrices and $\mathbf{R}_i$ and $\mathbf{R}_j$ are invertable upper-triangular matrices. The orthogonalized matrices $(\mathbf{W}^T \tilde{\mathbf{U}}_i)$ and $(\mathbf{W}^T \tilde{\mathbf{U}}_j)$ can be obtained as follows

$$\begin{aligned}
\Phi_i &= (\mathbf{W}^T \mathbf{U}_i) \cdot \mathbf{R}_i^{-1} = \mathbf{W}^T \cdot (\mathbf{U}_i \mathbf{R}_i^{-1}) = \mathbf{W}^T \tilde{\mathbf{U}}_i \\
\Phi_j &= (\mathbf{W}^T \mathbf{U}_j) \cdot \mathbf{R}_j^{-1} = \mathbf{W}^T \cdot (\mathbf{U}_j \mathbf{R}_j^{-1}) = \mathbf{W}^T \tilde{\mathbf{U}}_j
\end{aligned} \tag{6.7}$$

where

$$\tilde{\mathbf{U}}_i = \mathbf{U}_i \mathbf{R}_i^{-1}$$
$$\tilde{\mathbf{U}}_j = \mathbf{U}_j \mathbf{R}_j^{-1} \tag{6.8}$$

In Eq. (6.4) and Eq. (6.5), $tr(\Lambda)$ describes the correlation between two orthonormal basis matrices. Furthermore, $tr(\Lambda) \geq 0$ due to the fact that the principal angles are defined in the range of $[0, \frac{\pi}{2}]$. Since $\mathbf{U}_i$ and $\mathbf{U}_j$ are two different coordinate systems with respect to μ_i and μ_j respectively, the first term in Eq. (6.4) and Eq. (6.5) denotes how far away the origins of the two coordinate systems $\mathbf{U}_i$ and $\mathbf{U}_j$ are from each other in the subspace $\mathbf{W}$. Whereas the second term in Eq. (6.4) and Eq. (6.5) indicates that if the two coordinate systems are both shifted to the same origin, how relevant the two subspaces are. Based on Eq. (6.5), we observe that learning a good distance metric is equivalent to learn the linear transformations $\mathbf{W}$ for image sets in the original space.

6.2.3 Learning Parameter Matrix for the Distance Metric

With N image sets from C classes, N intrinsically low dimensional manifolds $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_N$ are embedded in the high dimensional ambient space. In order to give a more compact description of a manifold meanwhile avoid the curse of dimensionality, a nonlinear dimension reduction procedure [9–11] is applied on a manifold. Based on the proof [8] that the Laplacian of a graph can be considered as a good approximation of Laplace-Beltrami operator defined on the manifold, many graph embedding algorithms have been proposed to linearly map a manifold from a high dimensional Euclidean space to a low dimensional subspace, such that the embedding is defined everywhere in ambient space rather than just on the training data. However, these methods [8], [118] treated all the training samples from different classes with the same embedding and ignored that manifolds across classes might distribute very differently, i.e., there are large variations in the intrinsic structures across different classes. In view of this, we learn a class-specified distance metric $d_c^{\text{aff}}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{W}_c)$ for each class c. Then the proposed method aims to learn a collection of linear transformations $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_C$ from multiple manifolds.

There are N manifolds $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_N$ that constitute the whole training dataset $\mathbf{X} = [\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \cdots, \mathbf{x}_{1,l_1}, \cdots, \mathbf{x}_{N,1}, \mathbf{x}_{N,2}, \cdots, \mathbf{x}_{N,l_N}]$. For notation simplicity, $\mathbf{X}$ is denoted

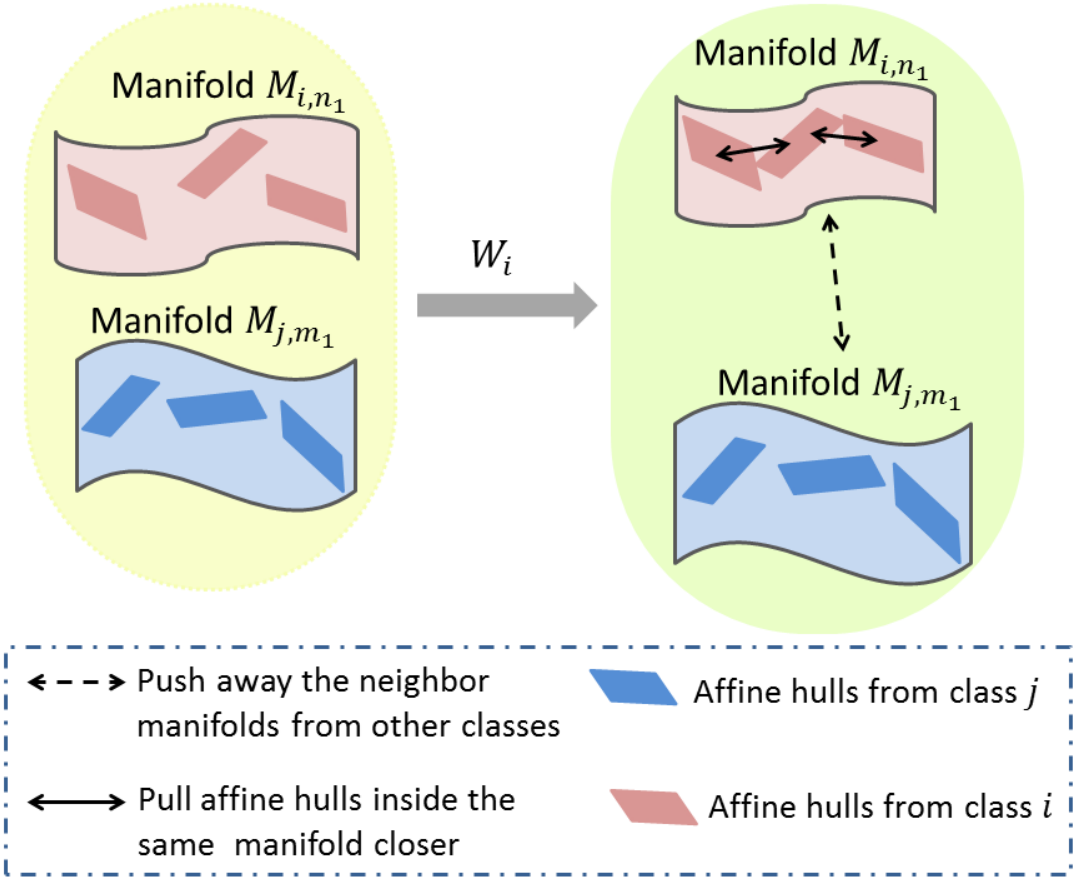

Figure 6.2: Illustration of the basic idea of the learning procedure for the ith class. For class i, we learn a distance metric that pushes the manifolds of class i away from neighbor manifolds of other classes meanwhile pulls affine hulls inside the same manifold closer.

as $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l]$ where $l = \sum_{i=1}^{N} l_i$. From the classification perspective, we expect that the learned transformations $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_C$ are able to maximize the scatterness between neighboring manifolds but from different classes, and minimize the compactness of manifolds from the same class simultaneously. Figure 6.2 shows the basic idea of the learning procedure for the ith class. To achieve this goal, we maximize the discriminability of training data by solving the following objective function for the cth class:

$$\max_{\mathbf{W}_c} f_b(\mathbf{W}_c), \quad \text{s.t.} \quad f_w(\mathbf{W}_c) = 1, \tag{6.9}$$

where

$$f_b(\mathbf{W}_c) = \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_b^{(c)}(i,j), \tag{6.10}$$

$$f_w(\mathbf{W}_c) = \sum_{i,j} \|\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j\|^2 G_w^{(c)}(i,j). \tag{6.11}$$

For the c^{th} class, the penalty graph $G_b^{(c)}$ that describes the between-class scatterness introduces a penalty of $||\mathbf{W}_c^T \mathbf{x}_i - \mathbf{W}_c^T \mathbf{x}_j||$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ is near to each other but from different classes, whereas the intrinsic graph $G_w^{(c)}$ encourages $\mathbf{W}_c^T \mathbf{x}_i$ and $\mathbf{W}_c^T \mathbf{x}_j$ to get closer if $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same class. The penalty graph $G_b^{(c)}$ and intrinsic graph $G_w^{(c)}$ are defined as:

$$G_b^{(c)}(i,j) = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|}{t_b}}, & \text{if } \mathbf{M}(\mathbf{x}_j) \in N_k(\mathbf{M}(\mathbf{x}_i)) \text{ ,} \\ & \text{where only } \mathbf{x}_i \in \text{class c,} \\ 0, & \text{otherwise,} \end{cases} \tag{6.12}$$

$$G_w^{(c)}(i,j) = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|}{t_w}}, & \text{if both } \mathbf{x}_i, \mathbf{x}_j \in \text{class c,} \\ 0, & \text{otherwise,} \end{cases} \tag{6.13}$$

where t_b and t_w are parameters of the heat kernels that follow an Gaussian distribution approximately in a local area. $\mathbf{M}(\mathbf{x}_i)$ denotes the manifold to which $\mathbf{x}_i$ belongs, and $N_k(\mathbf{M}(\mathbf{x}_i))$ denotes the manifolds that are the k nearest neighbors of manifold $\mathbf{M}(\mathbf{x}_i)$.

For ease of analysis, we simplify $f_b(\mathbf{W}_c)$ in (6.10) as:

$$\begin{aligned}
&\sum_{i,j} \|\mathbf{W}_c^T\mathbf{x}_i - \mathbf{W}_c^T\mathbf{x}_j\|^2 G_b^{(c)}(i,j) \\
= \;& tr\left\{\mathbf{W}_c^T\left[\sum_{i,j}(\mathbf{x}_i-\mathbf{x}_i)(\mathbf{x}_i-\mathbf{x}_j)^T G_b^{(c)}(i,j)\right]\mathbf{W}_c\right\} \\
= \;& tr\left\{\mathbf{W}_c^T\left[\sum_{i,j}\mathbf{x}_i G_b^{(c)}(i,j)\mathbf{x}_i^T - \sum_{i,j}\mathbf{x}_j G_b^{(c)}(i,j)\mathbf{x}_i^T\right.\right. \\
& \left.\left. - \sum_{i,j}\mathbf{x}_i G_b^{(c)}(i,j)\mathbf{x}_j^T + \sum_{i,j}\mathbf{x}_j G_b^{(c)}(i,j)\mathbf{x}_j^T\right]\mathbf{W}_c\right\} \\
= \;& tr\left\{\mathbf{W}_c^T\left[\sum_{i}\mathbf{x}_i D_b^{(c)}(i,i)\mathbf{x}_i^T - \mathbf{X}\mathbf{G}_b^c\mathbf{X}^T\right.\right. \\
& \left.\left. -\mathbf{X}\mathbf{G}_b^c\mathbf{X}^T + \sum_{j}\mathbf{x}_j D_b^{(c)}(j,j)\mathbf{x}_j^T\right]\mathbf{W}_c\right\} \\
= \;& 2tr\left(\mathbf{W}_c^T\mathbf{X}\mathbf{D}_b^{(c)}\mathbf{X}^T\mathbf{W}_c - \mathbf{W}_c^T\mathbf{X}\mathbf{G}_b^{(c)}\mathbf{X}^T\mathbf{W}_c\right) \\
= \;& 2tr\left(\mathbf{W}_c^T\mathbf{X}\mathbf{L}_b^{(c)}\mathbf{X}^T\mathbf{W}_c\right), \qquad (6.14)
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with $D_b^{(c)}(i,i) = \sum_j G_b^{(c)}(i,j)$ and $\mathbf{L}_b^{(c)} = \mathbf{D}_b^{(c)} - \mathbf{G}_b^{(c)}$ is the Laplacian of the penalty graph $\mathbf{G}_b^{(c)}$.

In the similar way, $f_w(\mathbf{W}_C)$ in (6.11) can be simplified as follows:

$$\begin{aligned}
&\sum_{i,j} \|\mathbf{W}_c^T\mathbf{x}_i - \mathbf{W}_c^T\mathbf{x}_j\|^2 G_w^{(c)}(i,j) \\
= \;& 2tr\left(\mathbf{W}_c^T\mathbf{X}\mathbf{L}_w^{(c)}\mathbf{X}^T\mathbf{W}_c\right), \qquad (6.15)
\end{aligned}$$

where the multiplier 2 is a constant that can be discarded. Thus, the optimal transformation matrix $\mathbf{W}_c$ can be derived by solving the generalized eigen-decomposition of matrices $\mathbf{L}_w^{(c)}$ and $\mathbf{L}_b^{(c)}$:

$$\mathbf{X}\mathbf{L}_b^{(c)}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{L}_w^{(c)}\mathbf{X}^T\mathbf{w}. \qquad (6.16)$$

The largest d eigenvectors $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d\}$ corresponding to the largest d eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ are chosen as the transformation matrix $\mathbf{W}_c$. In this way, in the learning phase we compute $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_C$ for each class respectively.

6.2.4 Recognition

Having obtained a collection of discriminant parameter matrix $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_C$, the recognition procedure is conducted on a probe image set, the manifold of which is denoted as $\mathbf{M}_p$. The probe set is first divided into m_p clusters by using a spectral clustering algorithm and then characterized as a set of affine hulls $\mathbf{M}_p = [\mathbf{h}_1^{(p)}, \mathbf{h}_2^{(p)}, \cdots, \mathbf{h}_{m_p}^{(p)}]$. Similarly, a gallery set from class c is also described as $\mathbf{M}_g^{(c)} = [\mathbf{h}_1^{(c,g)}, \mathbf{h}_2^{(c,g)}, \cdots, \mathbf{h}_{m_g}^{(c,g)}]$. To compare the probe manifold to the gallery manifold, we directly match them using the manifold to manifold distance metric whose parameter matrix $\mathbf{W}_c$ is obtained from learning:

$$d^{\text{mf}}(\mathbf{M}_p, \mathbf{M}_g) = \min_{\mathbf{h}_i^{(p)} \in \mathbf{M}_p} \min_{\mathbf{h}_j^{(c,g)} \in \mathbf{M}_q} d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c). \tag{6.17}$$

The probe manifold is assigned the label of a gallery manifold which is the nearest neighbor to the probe.

To speed up the proposed algorithm, we let $\alpha = 1$ in (6.5). Thus the affine hull distance metric becomes:

$$\begin{aligned} & d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c\mathbf{W}_c^T) \\ = & \frac{\mu_i^{(p)T}\mathbf{W}_c\mathbf{W}_c^T\mu_j^{(c,g)}}{(\mu_i^{(p)T}\mathbf{W}_c\mathbf{W}_c^T\mu_i^{(p)})(\mu_j^{(c,g)T}\mathbf{W}_c\mathbf{W}_c^T\mu_j^{(c,g)})}, \end{aligned} \tag{6.18}$$

where the mean vectors $\mu_j^{(c,g)}$ and $\mu_i^{(p)}$ are used as descriptors of affine hull $\mathbf{h}_j^{(c,g)}$ and $\mathbf{h}_i^{(p)}$ respectively. Under such circumstances, the Euclidean distance between two mean vectors $\mu_i^{(p)}$ and $\mu_j^{(c,g)}$ can be used as an alternative affine hull distance metric:

$$\begin{aligned} & d^{\text{aff}}(\mathbf{h}_i^{(p)}, \mathbf{h}_j^{(c,g)}, \mathbf{W}_c\mathbf{W}_c^T) \\ = & \sqrt{(\mu_i^{(p)} - \mu_j^{(c,g)})^T\mathbf{W}_c\mathbf{W}_c^T(\mu_i^{(p)} - \mu_j^{(c,g)})}. \end{aligned} \tag{6.19}$$

6.3 Experiments

In this section, we evaluate the proposed MMML method on the task of face recognition based on image sets. Given a query face image set, its nearest gallery set is found by using the collection of parameter matrices $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_c$ and the nearest neighbor classifier.

6.3.1 Experimental Setup

Goals: Through the following experiments, we aims to evaluate the performance of the proposed MMML method in several aspects to verify its efficacy.

- In experiment 1, in order to demonstrate the effectiveness of using affine hull models, we compare the performance of the methods based on the affine hull model to that based on the linear subspace model (PCA model).

- In experiment 2, in order to show the effectiveness of using multiple class-specific subspaces, we test the performance of the methods based on multiple class-specific subspaces and one unified subspace respectively.

- In experiment 3, we illustrate the effect of the number of affine hulls on the recognition performance of the proposed MMML method.

- In experiment 4, we show the effect of the number of dimensions of each class-specific subspace on the recognition performance of the proposed MMML method.

- In experiment 5, we compare the proposed MMML method with five state of the arts on three widely studied face video databases to demonstrate the effectiveness of the proposed method.

The three databases used for evaluations: We evaluate the performance of the proposed MMML method based on three widely studied face video databases: the Honda/UCSD database [127], the CMU MoBo database [119] and the Youtube Celebrities database [128]. Each video sequence in these databases is broken into frames and saved as an image set.

- **The Honda/UCSD Database**: As introduced in Section 4.4.1, Honda/UCSD database [127] contains 59 video clips involving 20 individuals. There are some exemplar cropped face images from the Honda/UCSD database are shown in Fig. 4.3.

- **The CMU MoBo Database**: As introduced in Chapter 3, there are 96 videos sequences from 24 different persons in the CMU MoBo (Motion of body) database [119] where each person has 4 video sequences containing a large number of variations in poses and expressions. Some exemplar face images from the CMU MoBo database are shown in Fig. 3.9.

- **The Youtube Celebrities Database**: As introduced in Section 4.4.1, the Youtube Celebrities Database [128] contains 1910 video sequences from 47 different individuals. These videos are obtained from the real-life scenarios and with low qualities. Some well-cropped face images from the Youtube Celebrities database are shown in Fig. 4.4.

The six approaches used for comparisons: In the experimental parts, we compare the proposed MMML method with the state of the arts to demonstrate its efficacy. These methods are: the Affine hull-based Image Set Distance (AHISD) [98], Discriminant Canonical Correlation Analysis (DCC) [104], Manifold-to-Manifold Distance(MMD) [16], Manifold Discriminant Analysis (MDA) [118] and Sparse Approximated Nearest Points(SANP) method [117]. Among these methods, there two unsupervised set-based matching methods including the MMD method and the AHISD method. The other three algorithms are based on supervised learning stages.

- **The DCC method:** A discriminative learning approach using canonical correlations known as DCC was proposed in reference [104]. Inspired by both the linear discriminant analysis (LDA) method and the canonical correlations analysis (CCA) method, the DCC algorithm defines the distance between image sets by using canonical correlations and uses a discriminative learning method to maximize the between-class correlations and minimize the within-class correlations over all the image sets. The objective function optimizes the orthonormal basis matrix, rotation matrices and discriminant transformation matrix in each step iteratively and the convergence of the DCC method is verified through experiments. The source codes of the DCC method are provided by original authors. In our implementations, the important parameters are tuned and optimized as follows. The

PCA algorithm is applied to preserve 90% energy of each subspace such that the dimension of each subspace is 10. The dimension of the embedding subspace is set to 150.

- **The MDA method:** Manifold Discriminant Analysis (MDA) was proposed in reference [118] to improve their previous work, the MMD method. Different from MMD that utilizes the MLP to construct each subset, the MDA method adopts a Hierarchical Divisive Clustering (HDC) method that is able to divide more balanced subsets than the MLP method. Then the MDA method finds a discriminative transformation for all the manifolds in the training dataset to maximize the separability of local models from different manifolds and minimize the compactness of images from the same local model. The source codes of the MDA method are provided by original authors. To achieve the optimal performance, we tune the important parameters in our implementation as follows: The ratio between geodesic distance and Euclidean distance is tuned to 1.1 to achieve an optimal performance. The other parameters are set following the experimental settings in their papers.

- **The SANP method:** The Sparse Approximated Nearest Points(SANP) method was recently proposed in the reference[117]. It utilizes an affine hull to describe each image set. This is different from our MMML method where an image set is described by a collection of affine hulls. In their method, the distance between two image sets is obtained from the nearest pair of points from two affine hulls, where the nearest point is approximated by the sparse linear combination of images in the other set. The sparse approximation error is used as the distance between two image sets. The source codes of the SANP algorithm are also provided by authors, which facilitates the implementations. We set the important parameters, all the weights for convex optimization, as the same the shown in their paper.

- **The MMD method:** As introduced in the previous chapters, the Manifold-to-manifold distance (MMD) method utilizes the MLP algorithm to cluster each image set and defines the manifold-to-manifold distance as the distance between the nearest pair of local models from two manifolds. In our implementations, we used the codes provided by the authors and set the parameters as follows: The ratio between

geodesic distance and Euclidean distance is tuned to 1.1 to achieve an optimal performance. The principal angles which are used in the manifold-to-manifold distance are the first 10 angles.

- **The AHISD method:** As introduced in Chapter 5, the Affine hull-based Image Set Distance (AHISD) method describes each image set by using a convex hull or an affine hull and the distance between two such convex geometric regions is defined as the distance between the nearest pair of points from two hulls. The classification is accomplished by a SVM classifier. The source code of the AHISD approaches is provided by original authors. In the implementations of the AHISD method, there is no parameter to be tuned.

- **The propsoed MMML method:** For the proposed MMML method, the parameter k of the k-means algorithm is set as follows: $k = 4$ if the number of images in each set is less than 20, $k = 9$ if the image number is less than 50. To speed up our algorithm, we adopt the affine hull-based distance metric defined in (6.18) and (6.19) where each affine hull is described by its mean vector.

6.3.2 Experiment 1: Affine Hull *vs* Linear Subspace

In the proposed MMML method, we adopt the affine hulls instead of linear subspaces to represent a nonlinear manifold. From the geometric viewpoint, in general an affine hull is a more compact representation than a linear subspace. Thus the affine hull-based methods are expected to perform better than the linear subspace-based methods in terms of the recognition performance. To verify the advantage of the proposed MMML method, we design experiment 1 as follows. Based on the Honda/UCSD database, we select 20 video sequences as gallery sets, with the remaining 39 videos as probe sets. For fair comparisons between the affine hull-based method and the linear subspace-based method, we adopt the same algorithm in the design of each step except the modeling part, where each subset is modeled using the affine hulls and linear subspaces respectively. It is noted that this experiment does not involve any learning stage for reducing computational complexity. Under such conditions, we randomly repeat the experiment 30 times. Then the averaged experimental results are given in Table 6.1 for the two methods respectively.

Table 6.1: Experimental results of experiment 1 on the Honda/UCSD database.

Methods	Average
Linear subspace-based method	$87.2\% \pm 4.1\%$
Affine hull-based method	$\mathbf{95.4}\% \pm 3.3\%$

In the Table. 6.1, it can be observed that the affine hull-based method improves the recognition performance significantly, which achieves better recognition rate 95.4% that outperforms that of the linear subspace-based method by around 8.2%.

6.3.3 Experiment 2: Multiple Class-specific Subspaces *vs* One Unified Subspace

Different from the existing methods that only learn a unified subspace for all the training samples, the proposed MMML method learns a collection of class-specific subspaces $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_C$ based on the observation that the manifolds in different classes distribute differently. To illustrate the efficacy of the proposed method using multiple class-specific subspaces, we design experiment 2 as follows. On the CMU MoBo database, one image set per person is randomly selected for training, with the remaining three image sets used for testing. For fair comparisons between the class-specific subspace-based method and the unified subspace-based method, we adopt the same algorithm in the design of each step except the learning part, where multiple class-specific subspaces and one unified subspace are learned respectively. Based on these settings, we repeat the experiment 30 times for each method. Then the averaged results of experiment 2 on the CMU MoBo database are illustrated in Table 6.2.

Table 6.2: Experimental results of experiment 2 on the CMU MoBo database.

Methods	Average
Unified based method	$95.8\% \pm 1.3\%$
Class-specific based method	$\mathbf{98.6}\% \pm 2.3\%$

We can observe from Table 6.2 that the recognition performance of the proposed method based on multiple class-specific subspaces reaches 98.6%, which significantly outperforms that of the method based on one unified subspace.

6.3.4 Experiment 3: Number of Affine Hulls

In the proposed MMML method, the spectral clustering algorithm is adopted to divide an image set into multiple clusters. The number of affine hulls, which we denote as k, equals to the number of clusters. To investigate how the selection of the number of affine hulls affects recognition performance, we conduct a series of experiments on Honda/UCSD database where 50 frames of each video are used. We test the MMML method with varying k (from 1 to 29) since the number of frames of each video varies from 29 to 50. For each k, we repeat the experiment 30 times and the averaged recognition rates are recorded. Thus we illustrate the relationship between the number of affine hulls and the recognition performance of the proposed method based on the Honda/UCSD database in Fig. 6.3.

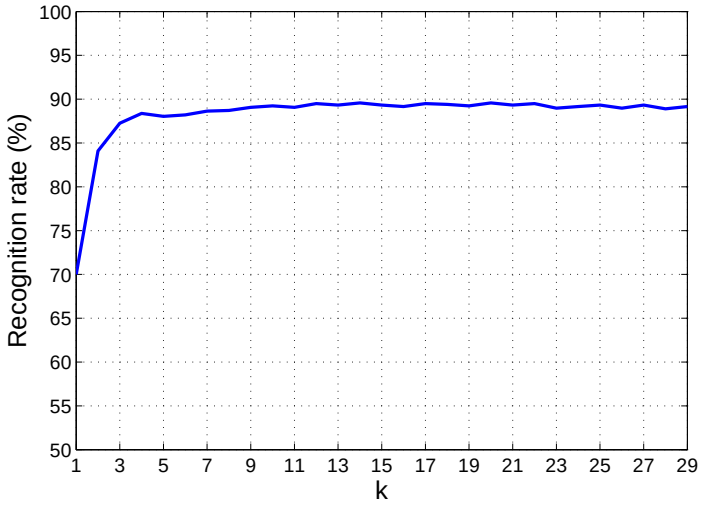

Figure 6.3: The recognition performance of the MMML method with varying number of affine hulls.

From the Fig. 6.3, it is observed that the performance of MMML method is robust with respect to varying k. When k=1, the recognition rate of MMML method only reaches 70% since the whole image set is directly put into the discriminative learning stage without division. With the increasing value of k, the recognition rate of the MMML method first rises and then stabilizes around 89% as shown in Fig. 6.3. Thus when k varies from 4 to 29, it is apparent that the proposed MMML method is not sensitive to

k. Hence, there is a large space for us to choose a suitable number of affine hulls for the proposed MMML method.

6.3.5 Experiment 4: The Dimension of Each Subspace

To further illustrate the robustness of the proposed MMML method, we evaluate its performance with an increase in the dimensions of each subspace in experiment 4. Similar to experiment 2, 96 face image sets from the CMU MoBo database are used. One image set per individual is randomly selected for learning, with the remaining three image sets per person used as the probe sets. We repeat the random selection 30 times for each subspace of fixed dimensions, and then increase the dimensions with step size 1. Then the averaged recognition rates are shown in Fig. 6.4 and 6.5, where the relationship between the number of dimensions of each discriminative subspace and the recognition rates of the proposed method on the CMU MoBo database is demonstrated. Fig. 6.4 shows the recognition rates when the number of dimensions varies from 1 to 60. To better illustrate the effect of the dimension on the performance of the proposed method, Fig. 6.5 demonstrates the recognition rates when the number of dimensions varies from 1 to 12.

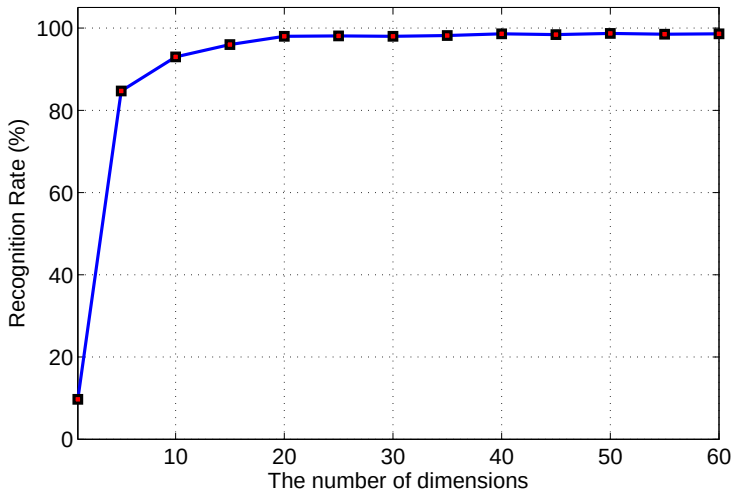

Figure 6.4: The recognition performance of the MMML method with increasing number of dimensions (from 1 to 60) of each discriminative subspace.

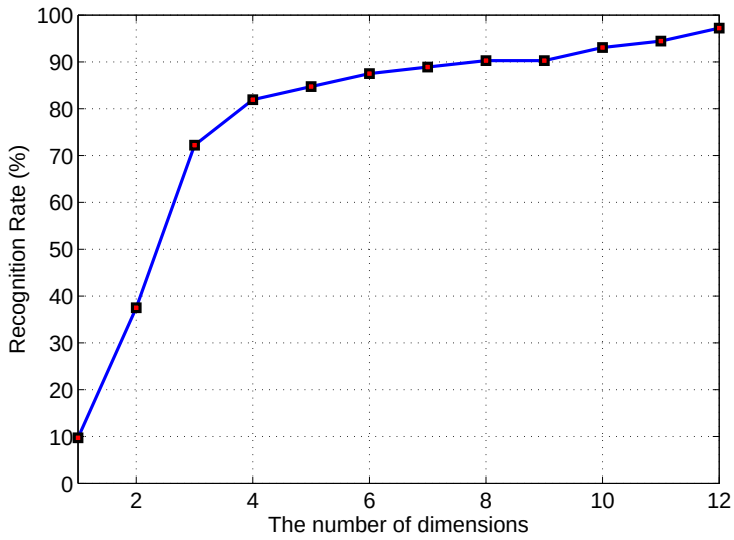

Figure 6.5: The recognition performance of the MMML method with increasing number of dimensions (from 1 to 12) of each discriminative subspace.

From Fig. 6.4 and 6.5, we observe that the performance of the proposed MMML method improves with an increase in the number of dimensions of each discriminative subspace. In Fig. 6.5, the recognition rate is only around 9.72% when the number of dimensions equals to one and quickly rises up to 37.5%, 72.22%, 81.94% and 84.72% with dimensions 2, 3, 4 and 5 respectively. However, when the number of dimensions keeps increasing beyond 12, the improvement in the recognition performance of the proposed MMML method is limited as shown in Fig. 6.4, which demonstrates a robustness to the dimension of each subspace. This provides us a large space to choose a suitable number of dimensions of each subspace for the proposed MMML method.

6.3.6 Experiment 5: Compare with Five State of the Arts

We compare the proposed MMML method with five set-based classification methods, the Affine hull-based Image Set Distance (AHISD) [98], Discriminant Canonical Correlation Analysis (DCC) [104], Manifold-to-Manifold Distance (MMD) [16], Manifold Discriminant Analysis (MDA) [118] and Sparse Approximated Nearest Points (SANP) method [117], as introduced in the experimental setting part.

Table 6.3: Experimental results on the Honda/UCSD database.

Methods	50 Frames	100 Frames	Full Length	Average
DCC	79.5%	**92.3**%	94.9%	88.9%
MMD	82.1%	84.6%	94.9%	87.2%
MDA	76.9%	89.7%	**100**%	88.9%
AHISD	87.2%	84.6%	89.7%	87.2%
SANP	84.6%	**92.3**%	**100**%	92.3%
MMML	**89.7**%	**92.3**%	97.22%	**93.1**%

Experiments on the Honda/UCSD Database: For the Honda/UCSD database [127], each cropped face image is resized to 20×20 as in [117] [16]. To minimize the effect of illuminations, the histogram equalization is performed as a preprocessing step.

In our implementation, we adopt the configuration described in Section 4.4.2 where 20 video sequences are selected for training and the remaining 39 videos are used for testing. Similar to [117], we evaluate the proposed method and the five methods under three settings: 50 frames each video sequence, 100 frames each video sequence and full length of each video sequence. The number of k nearest neighbor affine hulls is set to 7. The experimental results of the proposed method and the competing methods are summarized and shown in Table 6.3.

From the comparison results shown in Table 6.3, the proposed MMML method achieves the best average recognition rate 93.1%. Since the MMD and AHISD methods directly compare the models of image sets without any learning procedure, they perform not as well as the discriminative learning approaches DCC, MDA and MMML methods. Although the SANP does not involve any discriminative learning steps, it reconstructs a sample point using sparse affine coefficients which introduce the discriminative information into the sparse representations. Thus the SANP method is also able to achieve a comparable recognition rate. In the case that the frames of a video sequence are reduced, i.e., 50 frames in each image set, the performances of MDA and DCC degrade heavily because of the training samples is not enough for discriminative learning. In this situation, since the MMML method describes an image set using a collection of affine hulls which can "fill-in" the missing data, the MMML method shows effectiveness and robustness across different training set size, even the training samples decreased.

Table 6.4: Experimental results on CMU MoBo database.

Methods	Average
DCC	$94.9\% \pm 2.3\%$
MMD	$94.4\% \pm 3.0\%$
MDA	$95.8\% \pm 1.9\%$
AHISD	$94.4\% \pm 1.9\%$
SANP	$97.8\% \pm 1.1\%$
MMML	$\mathbf{98.6}\% \pm 2.5\%$

Experiments on CMU MoBo Database: For the CMU MoBo database, the cropped face images are resized to the size of 40×40 and the LBP features of each face image are extracted as described in the previous chapters. The configuration of experiments on the CMU MoBo datasets is: one image set per person is randomly selected for training and the remaining three image sets are used for testing. We repeat the random selection 30 times for each algorithm and the averaged results are illustrated in Table 6.4.

From the averaged recognition rates shown in Table 6.4, we observe that the performance of the proposed MMML method achieves 98.61% which shows the best average recognition rate over all methods. Compared to the Honda/UCSD database, the CMU-mobo database is more noisy, thus the proposed MMML method perform better than others since it learns a class-specific metric which is able to describe a noisy image set better. The recognition rate of SANP method reaches 97.8% which is slightly inferior to MMML. However the computational complexity of the proposed MMML method is much lower than SANP. On the CMU MoBo database, MMD and AHISD still perform not as well as discriminative learning methods such as MDA, DCC, SANP and MMML, where SANP employs sparse coefficients which introduce discriminative information.

From the Table 6.3 and Table 6.4, all the six set-based approaches perform well on the Honda/UCSD database and the CMU MoBo database, since the two databases contain the variations in pose and expression under controlled conditions where there is little variations in illumination and the videos are captured in the indoor environment.

Experiments on YouTube Celebrities Database: For the Youtube Celebrities database, each cropped face image is histogram equalized and resized to 50×50.

Table 6.5: Experimental results on YouTube Celebrities database.

Methods	Average
DCC	$62.2\% \pm 2.1\%$
MMD	$63.6\% \pm 2.3\%$
MDA	$61.9\% \pm 3.4\%$
AHISD	$64.2\% \pm 4.1\%$
SANP	$63.9\% \pm 2.3\%$
MMML	$\mathbf{69.4}\% \pm 2.8\%$

In our implementations, we follow the experimental configuration mentioned in [117] where there are five-fold cross validation experiments conducted on this database. We divide video sequences of each person into 5 groups and make sure that each group contains 9 videos(with minimum overlapping). After this division, the YouTube Celebrities Database becomes five folds. Each fold consists of 423 video sequences from 47 subjects. In each fold, we randomly select 3 image sets per person used in the learning stage and use the remaining 6 image sets in the classification phase. We repeated the random selected experiments 10 times for each algorithm and show the averaged recognition rates and standard deviations in Table 6.5.

In Table 6.5, all the set-based image classification approaches show lower performances on the YouTube Celebrities database, since the videos are obtained from real life with low quality, misalignments and the large variations in appearances, poses and expressions. In this challenging case, the recognition rate of our MMML method achieves 69.4% which outperforms all the other competing methods. Due to the single affine hull model cannot exactly describe an noisy image set which is quality-low with full of large appearance variations, the performance of the SANP method is not as well as that on Honda/CUCSD and CMU MoBo databases. Both the MAD and DCC methods learn a single discriminative subspace for all the samples from different classes. However the large variations cannot be covered into this single subspace which leads to the relative low recognition rates of MDA and DCC tested on the YouTube Celebrities database. In our method, the proposed affine hull-based distance metric is designed for each class such that the variations of the current class is characterized well by the class-specified model. This leads to the robust performance of our MMML method across the Honda/UCSD, CMU MoBo and YouTube Celebrities databases.

Table 6.6: Computation time of different methods on Honda/UCSD database (classification of one image set).

Methods	Training	Testing
DCC	$6.50s$	$0.23s$
MMD	N/A	$30.10s$
MDA	$8.90s$	$0.18s$
AHISD	N/A	$5.20s$
SANP	N/A	$2.80s$
MMML	$60.08s$	$0.13s$

6.3.7 Discussion

Computational Complexity: Table 6.6 compares the computational time of different methods on Honda/UCSD database. From this table, we can see that MMD spends the longest time 30.1 seconds in the testing step of classifying one image set, because two parts of MMD are time-consuming: constructing a collection of linear subspaces through PCA and matching image sets by computing their principal angles. From the recognition rates shown in Tables 6.3 and 6.4, we observe that the performances of SANP and MMML keep close to one another, but SANP costs more time (2.8 seconds) in classifying one image set than our MMML method (0.13 seconds). Although our proposed MMML needs to learn a distance metric for each class, the time consuming training step can be finished offline such that the testing time of classifying each probe image set does not increase.

Sample Size: The whole algorithm assumes that there are sufficient images in each image set such that a collection of affine hulls can be utilized to characterize a nonlinearly distributed manifold. Naturally, the small sample size (SSS) problem may arise in some scenarios. When there are not sufficient samples in an image set, the number of affine hulls used to describe this manifold will be reduced. Furthermore, for each subset of a manifold, there are two key factors to effectively describe its affine hull representation: the mean vector μ and the orthonormal basis matrix $\mathbf{W}$. When there are not sufficient samples in a subset, the mean vector is able to describe the subset better than the orthonormal basis matrix. Thus we give a larger weight to the mean vector and a smaller weight to the orthonormal basis matrix, such that our approach is able to adapt to different image sizes, and vice versa.

6.4 Conclusion

Different from the previous chapters that studied the approaches in a unsupervised manner, we conducted our investigation on the area of supervised multi-model based methods for face recognition with image sets in Chapter 6. Based on the motivation mentioned in Section 6.1, we discussed the proposed multi-manifold metric learning (MMML) method for face recognition from image sets in Section 6.2. We considered each image set as a nonlinear manifold and characterized it by using a collection of affine hulls. We learned multiple person-specific distance metrics based on the training manifolds to maximize the separability of neighbor manifolds from different classes and minimize the compactness of manifolds from the same class, simultaneously, such that more discriminative information can be exploited for classification. Experimental results on three popular databases are presented to show the effectiveness and robustness of our proposed method in Section 6.3, followed by the conclusion in Section 6.4.

Chapter 7

Conclusion and Future Research

7.1 Conclusion

As a challenging problem in the area of pattern recognition, face recognition has received a large amount of research interest over the past several decades. Although great progress has been made on this topic and the techniques have been considered rather mature, face recognition systems are far from achieving desirable performance under uncontrolled conditions that usually occur in our real lives. One possible solution to reduce the degree of challenges is to utilize more face images of each individual, e.g., a collection of photos from the personal galley or frames from a video clip, which include more noises such as misalignments, occlusions and large variations in illumination, expression, pose, resolution, etc.

In this thesis, we have investigated the topic of face recognition based on image sets (FRBIS) where a set of face images of an individual can be collected from various social media sharing websites such as Facebook, YouTube, Flickr, etc. Based on the investigation, we further proposed some possible solutions to address this problem. In the following sections, we summarize our work in this thesis and sketch the further research directions.

7.1.1 Single Model-based Methods for FRBIS

As reviewed in Chapter 2, the literatures on the area of FRBIS can generally be divided into three categories: model-free methods, single-model based methods and multi-model

based methods, of which we investigated the two predominant categories, i.e., the single-model based methods and multi-model based methods.

Among the single model-based methods, we investigated the correlation-based approaches that are efficient for set-to-set matching in Chapter 3. The correlation-based approaches usually model an image set by using a linear subspace or a weighted subspace. Based on the study of correlation-based methods, we proposed a generalized subspace distance (GSD) framework to illustrate the underlying relationships among several state of the arts. Moreover, we observed that the weight assigned to each basis is often a fixed value, e.g., the corresponding eigenvalue. To improve the recognition performance, we introduced a parameter, which is learned from a discriminative learning step, into the proposed GSD framework. This leads to the proposed FOWSD and EWSD that have stronger ability to classify image sets from different classes. Furthermore, we extended the FOWSD and EWSD methods to their affine versions such that each image set can be fitted better by using an affine hull. We verified the efficacy of the proposed methods on four widely used databases for both the face recognition and object recognition tasks.

7.1.2 Multi-model based Methods for FRBIS — Local Model Construction

Although we have strived to explore the potential of the single-model based methods in Chapter 3, the performance of them seems to confront a bottleneck, e.g., the best recognition results appear to be stable around 93% on the CMU MoBo database. This phenomenon is possibly caused by the fact that the single-model based methods have the limited ability to describe the nonlinearly distributed image sets that are usually obtained from the real-life applications. In view of this observation, we exploited the multi-model based methods, which are more flexible to capture the nonlinear distributed image sets, for solving the FRBIS problem. Generally, the multi-model based methods consists of two steps: the local model construction and the manifold-to-manifold distance measure.

In Chapter 4, we investigated the first stage: local model construction. In the existing multi-model based methods, an image set is usually considered as a nonlinear manifold that is represented by using a collection of local models. Usually, a local model is a linear

subspace in the state of the arts. However, constructing a linear subspace from an image set and matching different linear subspaces are very time-consuming steps. In view of this, we adopted the mean vector of each subset as the local model to accelerate the procedure of computing manifold-to-manifold distances. The question is how to divide an image set into clusters such that the mean vector of each subset becomes more representative. The conventional clustering methods, the k-means clustering and standard spectral clustering methods, cluster images by using only one view of the image set. Different from this, we proposed a co-learned multi-view spectral clustering (CMSC) algorithm that combines multiple representations of an image set, e.g., the SIFT descriptors, LBP descriptors and intensity values, such that the resultant clusters are able to provide more representative mean vectors as local models. We verified the efficacy of the proposed CMSC method through a series of experiments.

7.1.3 Multi-model based Methods for FRBIS — Manifold to Manifold Distance

As a continuous of Chapter 4, we further investigated the stage of manifold-to-manifold distance measure for multi-model based approaches. The existing method for measuring the similarity between manifolds is to compare the most similar pair of local models from two manifolds respectively, which is denoted as nearest neighbor based manifold-to-manifold distance (NN-MMD). However, the NN-MMD cannot guarantee a reliable recognition result when an face image set contains outliers. To make the recognition system more robust to noises, we proposed a collaborative reconstruction-based manifold-to-manifold distance (CRMMD) that can be obtained from the minimal reconstruction error between a local model and its approximation, which is collaboratively reconstructed from several local models on the other manifold. Hence, the proposed CRMMD reduces its sensitivity to random variations in local models and provides more accurate classification results, which was verified through comparing with state of the arts on three widely studied databases.

7.1.4 Multi-model based Methods for FRBIS — Supervised Learning

From Chapter 4 and Chapter 5, we observed that the multi-model based methods which we had investigated are based on unsupervised matching stage. It is known that the supervised learning is very useful for pattern classification task, since the discriminative information provided by class labels are fully utilized in the training phase. Thus we attempted to exploit the discriminative learning methods to enhance the classification performance of the multi-model based methods. It is noted that the existing discriminative learning algorithm used for multi-model based methods only learns one single low-dimensional subspace for all the image sets from different classes where the image sets distributes differently. In contrast, we proposed a person-specific metric learning approach denoted as the multi-manifold metric learning (MMML) method for FRBIS in Chapter 6. Firstly, our MMML method adopted a metric learning manner. Different from the existing metric learning methods that learn the distance metric from single images and cannot deal with image sets, the proposed MMML method learns the distance metrics for matching manifolds. Secondly, the MMML method learns a collection of person-specific distance metrics, one for each class. This provides more discriminative distance metrics for classifying manifolds. Based on the above advantages, the proposed MMML algorithm has outperformed the state of the arts as shown in the experimental section.

7.2 Future Work

7.2.1 Possible Directions of FRBIS

Although we have investigated the problem of face recognition based image sets as completely as possible, there are still many potential solutions worth to be further explored.

For the FRBIS problem, our study is from the single model-based methods to the multi-model based methods and the experimental results have showed that the multi-model based methods generally outperform the single model-based methods. In view of this, one can continue to investigate the multi-model based methods. In our proposed methods introduced in the previous chapters, we used a mean vector to describe each

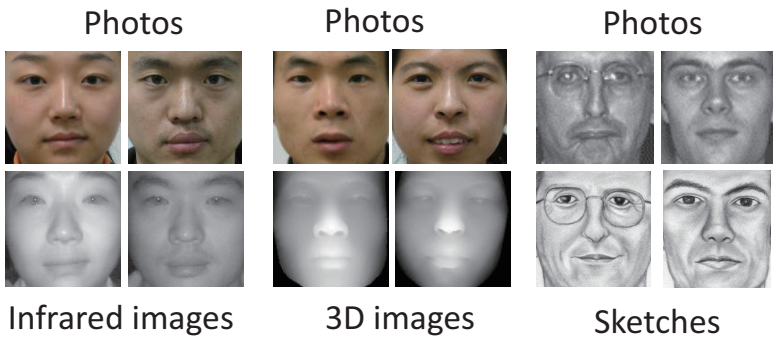

Figure 7.1: Three examples of heterogeneous face recognition applications. Form left to right are visual images v.s. NIR images, 2D images v.s. 3D images, and photo-sketch heterogeneous face recognition, respectively.

subset of an image set. However, the mean vector might have limited capability to capture a subset when this subset contains too many samples or distributes sparsely. Under this situation, it is better to adopt some more complex local models such as the covariance matrix, the low-rank matrix, etc. Accordingly, a new manifold-to-manifold distance measure should be re-designed to matching such complex local models.

Furthermore, the proposed CMSC method can be applied to more clustering applications and the proposed CRMMD methods can be applied to other set-based visual recognition applications such as video-based object search and multi-view object recognition to further demonstrate its effectiveness.

7.2.2 Heterogeneous Face Recognition

We have investigated the topic of face recognition based on image sets that is one of the problems confronted in the practical face recognition applications. Besides the FRBIS problem, there are still many practical problems of face recognition left and needed to be explored. For example, human can recognize faces by using information from multiple stimuli (the senses of eye, nose, ear and touch) and some experience such as race, gender and the relevance between individuals and contextual knowledge, etc. Inspired by this, more and more research work has looked into the areas of multi-model (combining the information of speech, visual and gait, etc.) face recognition or multi-modality (infrared face image, 3D face image and sketch, etc.) face recognition.

A challenging task is the heterogeneous face recognition where the query face images captured on spot and the reference face images stored in the database are captured from different modalities, e.g., visual images vs. near infrared (NIR) images, digital images vs. sketches, or 2D images vs. 3D images, as the examples shown in Fig. 7.1. Heterogeneous face recognition is a very common situation in our daily lives, for example:

- Visual images vs. near infrared (NIR) images: NIR is an emerging imaging modality which is able to combat the weak lighting condition and capture clear images of a subject in the night, cloudy or rainy days. In a heterogeneous face recognition system, the NIR images captured by surveillance camera is compared to optical images stored in the gallery database.

- Photos vs. sketches: Sketches drawn by artists are important tools for determining the identity of a suspect whose photo is unavailable when criminal activities occur.

- 2D images vs. 3D images: 3D face images has the potential to improve face recognition performance under the conditions with large variations in pose. The system often matching 3D face images enrolled in the gallery database with 2D probe images since the 2D images are not easy to obtain in many applications.

The difference in modalities caused by different acquisition processes leads to unsatisfactory face recognition results. Heterogeneous face recognition algorithms are designed to reduce the gap between different modalities and utilize the face images from various modalities.

Publication

Conference Papers

(i) Likun Huang, Jiwen Lu, Yap Peng Tan and Xin Feng, "Collaborative reconstruction-based manifold-manifold distance for face recognition with image sets." *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

(ii) Likun Huang, Jiwen Lu, and Yap-Peng Tan, "Learning modality-invariant features for heterogeneous face recognition." *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pp. 1683-1686. 2012. (oral presentation)

(iii) Likun Huang, Jiwen Lu, Gao Yang, and Yap-Peng Tan, "Generalized subspace distance for set-to-set image classification." *Proceedings of the 2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1123-1126. 2012. (oral presentation)

(iv) Likun Huang, Jianchao Yao, and Yap-peng Tan, "Enhancing incremental learning/recognition via efficient neighborhood estimation." *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 637-642. 2010.

Journal Papers

(i) Likun Huang, Jiwen Lu, and Yap-Peng Tan, "Co-Learned Multi-View Spectral Clustering for Face Recognition Based on Image Sets." *IEEE Signal Processing Letters*, vol. 21, No. 7, pp. 875–879 July 2014.

(ii) Likun Huang, Jiwen Lu, and Yap-Peng Tan, "Multi-Manifold Metric Learning for Face Recognition Based on Image Sets." *Accepted by the Journal of Visual Communication and Image Representation.*

References

[1] E. McKone, N. Kanwisher, B. C. Duchaine, *et al.*, "Can generic expertise explain special processing for faces?," *Trends in cognitive sciences*, vol. 11, no. 1, pp. 8–15, 2007.

[2] S. Carey, R. Diamond, *et al.*, "From piecemeal to configurational representation of faces," *Science*, vol. 195, no. 4275, pp. 312–314, 1977.

[3] S. Carey and R. Diamond, "Are faces perceived as configurations more by adults than by children?," *Visual Cognition*, vol. 1, no. 2-3, pp. 253–274, 1994.

[4] M. D. Kelly, "Visual identification of people by computer.," tech. rep., DTIC Document, 1970.

[5] T. Kanade, *Computer recognition of human faces*, vol. 47. Birkhäuser, 1977.

[6] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, pp. 71–86, 1991.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on PAMI*, pp. 711–720, 1997.

[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on PAMI*, pp. 328–340, 2005.

[9] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, pp. 2319–2323, 2000.

[10] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, pp. 2323–2326, 2000.

[11] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *NIPS*, pp. 585–591, 2001.

[12] O. Arandjelović and R. Cipolla, "An information-theoretic approach to face recognition from face motion manifolds," *Image and Vision Computing*, vol. 24, no. 6, pp. 639–647, 2006.

[13] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, pp. 582–585, IEEE, 1994.

[14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

[16] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[17] K. Kim, "Intelligent immigration control system by using passport recognition and face verification," in *Advances in Neural Networks–ISNN 2005*, pp. 147–156, Springer, 2005.

[18] J. N. Liu, M. Wang, and B. Feng, "ibotguard: an internet-based intelligent robot security system using invariant face recognition against intruder," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 1, pp. 97–105, 2005.

[19] H. Moon, "Biometrics person authentication using projection-based face recognition system in verification scenario," in *Biometric Authentication*, pp. 207–213, Springer, 2004.

[20] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings, International Conference on Audio-and Video-Based Person Authentication*, pp. 176–181, Citeseer, 1999.

[21] S. L. Wijaya, M. Savvides, and B. Vijaya Kumar, "Illumination-tolerant face verification of low-bit-rate jpeg2000 wavelet images with advanced correlation filters for handheld devices," *Applied optics*, vol. 44, no. 5, pp. 655–665, 2005.

[22] D. McCullagh, "Call it super bowl face scan i," *Wired, Feb*, vol. 2, 2001.

[23] K. Balci and V. Atalay, "Pca for gender estimation: which eigenvectors contribute?," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, pp. 363–366, IEEE, 2002.

[24] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 707–711, 2002.

[25] B. Poggio, R. Brunelli, and T. Poggio, "Hyberbf networks for gender classification," 1992.

[26] A. Colmenarez, B. Frey, and T. S. Huang, "A probabilistic framework for embedded face and facial expression recognition," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1, IEEE, 1999.

[27] Y. Shinohara and N. Otsuf, "Facial expression recognition using fisher weight maps," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 499–504, IEEE, 2004.

[28] F. Bourel, C. C. Chibelushi, and A. A. Low, "Robust facial feature tracking," in *British Machine Vision Conference*, vol. 1, pp. 232–241, 2000.

[29] K. Morik, P. Brockhausen, and T. Joachims, "Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pp. 268–277, MORGAN KAUFMANN PUBLISHERS, INC., 1999.

[30] D. Metaxas, S. Venkataraman, and C. Vogler, "Image-based stress recognition using a model-based dynamic face tracking system," in *Computational Science-ICCS 2004*, pp. 813–821, Springer, 2004.

[31] C. Nastar and M. Mitschke, "Real-time face recognition using feature combination," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 312–317, IEEE, 1998.

[32] S. Gong, S. J. McKenna, and A. Psarrou, *Dynamic vision*. Imperial College Press, 2000.

[33] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *journal of information processing systems*, vol. 5, no. 2, pp. 41–68, 2009.

[34] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, 2000.

[35] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[36] D. M. Blackburn, P. J. Phillips, and M. Bone, *Facial recognition vendor test 2000 evaluation report*. US Department of Defense, 2001.

[37] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, p. 44, IEEE, 2003.

[38] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, N. Poh, *et al.*, "Face authentication test on the banca database," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, pp. 523–532, IEEE, 2004.

[39] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, vol. 1, pp. 947–954, IEEE, 2005.

[40] P. J. Phillips, W. T. Scruggs, A. J. OToole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Frvt 2006 and ice 2006 large-scale results," 2007.

[41] Y. Gang, L. Jiawei, L. Jiayu, M. Qingli, and Y. Ming, "Illumination variation in face recognition: A review," in *Intelligent Networks and Intelligent Systems, 2009. ICINIS'09. Second International Conference on*, pp. 309–311, IEEE, 2009.

[42] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[43] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[44] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[45] S.-H. Lin, S.-Y. Kung, and L.-J. Lin, "Face recognition/detection by probabilistic decision-based neural network," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 114–132, 1997.

[46] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 6, pp. 748–763, 2002.

[47] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[48] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, *et al.*, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[49] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 681–685, 2001.

[50] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003.

[51] T. Kanade, "Picture processing system by computer complex and recognition of human faces," 1974.

[52] M. Nixon, "Eye spacing measurement for facial recognition," in *29th Annual Technical Symposium*, pp. 279–285, International Society for Optics and Photonics, 1985.

[53] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International journal of computer vision*, vol. 8, no. 2, pp. 99–111, 1992.

[54] N. Roeder and X. Li, "Experiments in analyzing the accuracy of facial feature detection," in *Vision Interface*, vol. 95, pp. 8–16, 1995.

[55] C. Colombo, A. Del Bimbo, and S. De Magistris, "Human-computer interaction based on eye movement tracking," in *Computer Architectures for Machine Perception, 1995. Proceedings. CAMP'95*, pp. 258–263, IEEE, 1995.

[56] I. J. Cox, J. Ghosn, and P. N. Yianilos, "Feature-based face recognition using mixture-distance," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pp. 209–216, IEEE, 1996.

[57] F. Samaria and F. Fallside, *Face identification and feature extraction using hidden markov models*. Citeseer, 1993.

[58] F. S. Samaria, *Face recognition using hidden Markov models*. PhD thesis, University of Cambridge, 1994.

[59] F. Samaria and S. Young, "Hmm-based architecture for face identification," *Image and vision computing*, vol. 12, no. 8, pp. 537–543, 1994.

[60] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp. 138–142, IEEE, 1994.

[61] A. V. Nefian and M. H. Hayes III, "Hidden markov models for face recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 5, pp. 2721–2724, IEEE, 1998.

[62] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 775–779, 1997.

[63] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The bochum/usc face recognition system and how it fared in the feret phase iii test," in *Face Recognition*, pp. 186–205, Springer, 1998.

[64] J. Buhmann, M. Lades, and C. von der Malsburg, "Size and distortion invariant object recognition by hierarchical graph matching," in *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pp. 411–416, IEEE, 1990.

[65] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *Computers, IEEE Transactions on*, vol. 42, no. 3, pp. 300–311, 1993.

[66] P. J. Phillips, P. J. Rauss, and S. Z. Der, *FERET (face recognition technology) recognition algorithm development and test results*. Defense Technical Information Center, 1996.

[67] I. Biederman and P. Kalocsai, "Neural and psychophysical analysis of object and face recognition," in *Face Recognition*, pp. 3–25, Springer, 1998.

[68] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, 1997.

[69] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 10, pp. 1042–1052, 1993.

[70] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[71] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990.

[72] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *JOSA A*, vol. 4, no. 3, pp. 519–524, 1987.

[73] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, no. 1, pp. 103–108, 1990.

[74] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591, IEEE, 1991.

[75] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[76] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 696–710, 1997.

[77] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[78] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *Neural Networks, IEEE Transactions on*, vol. 13, no. 6, pp. 1450–1464, 2002.

[79] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[80] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[81] A. M. Martinez and A. C. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228–233, 2001.

[82] J. R. Beveridge, K. She, B. A. Draper, and G. H. Givens, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–535, IEEE, 2001.

[83] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[84] W. Liu, Y. Wang, S. Z. Li, and T. Tan, "Null space approach of fisher discriminant analysis for face recognition," in *Biometric Authentication*, pp. 32–44, Springer, 2004.

[85] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[86] M. Li and B. Yuan, "2d-lda: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[87] H. Xiong, M. Swamy, and M. Ahmad, "Two-dimensional fld for face recognition," *Pattern Recognition*, vol. 38, no. 7, pp. 1121–1124, 2005.

[88] D. Zhou and X. Yang, "Face recognition using enhanced fisher linear discriminant model with facial combined feature," in *PRICAI 2004: Trends in Artificial Intelligence*, pp. 769–777, Springer, 2004.

[89] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, pp. 995–1006, 2004.

[90] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, pp. 982–994, 2004.

[91] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[92] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, pp. 585–591, 2001.

[93] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[94] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[95] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[96] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 40–51, 2007.

[97] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[98] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2567–2573, IEEE, 2010.

[99] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, 2006.

[100] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 581–588, IEEE, 2005.

[101] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Computer VisionECCV 2002*, pp. 851–865, Springer, 2002.

[102] L. Wang, X. Wang, and J. Feng, "Subspace distance analysis with application to adaptive bayesian algorithm for face recognition," *Pattern recognition*, vol. 39, no. 3, pp. 456–464, 2006.

[103] F. Li, Q. Dai, W. Xu, and G. Er, "Weighted subspace distance and its applications to object recognition and retrieval with image sets," *Signal Processing Letters, IEEE*, vol. 16, no. 3, pp. 227–230, 2009.

[104] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.

[105] O. Yamaguchi, K. Fukui, and K.-i. Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 318–323, IEEE, 1998.

[106] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Robotics Research*, pp. 192–201, Springer, 2005.

[107] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *The Journal of Machine Learning Research*, vol. 4, pp. 913–931, 2003.

[108] T. Kozakaya, O. Yamaguchi, and K. Fukui, "Development and evaluation of face recognition system using constrained mutual subspace method," *IPSJ J*, vol. 45, no. 3, pp. 951–959, 2004.

[109] k. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

[110] J.-M. Chang, M. Kirby, and C. Peterson, "Set-to-set face recognition under variations in pose and illumination," in *Biometrics Symposium, 2007*, pp. 1–6, IEEE, 2007.

[111] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 376–383, ACM, 2008.

[112] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[113] X. Sun, L. Wang, and J. Feng, "Further results on the subspace distance," *Pattern recognition*, vol. 40, no. 1, pp. 328–329, 2007.

[114] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 163–168, IEEE, 2000.

[115] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley-interscience, 2012.

[116] V. Vapnik, *The nature of statistical learning theory*. springer, 1999.

[117] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 121–128, IEEE, 2011.

[118] R. Wang and X. Chen, "Manifold discriminant analysis," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 429–436, IEEE, 2009.

[119] R. Gross and J. Shi, "The cmu motion of body (mobo) database," 2001.

[120] J.-M. Geusebroek, G. J. Burghouts, and A. W. Smeulders, "The amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[121] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[122] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–409, IEEE, 2003.

[123] A. Kumar, P. Rai, and H. Daumé III, "Co-regularized multi-view spectral clustering," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1413–1421, 2011.

[124] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, p. 14, California, USA, 1967.

[125] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[126] A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[127] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *CVPR*, pp. I–313, 2003.

[128] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *CVPR*, pp. 1–8, 2008.

[129] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, pp. 125–141, 2008.

[130] J. Hartigan, *Clustering algorithms*. 1975.

[131] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *NIPS*, pp. 505–512, 2002.

[132] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *NIPS*, 2005.

[133] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, pp. 207–244, 2009.

[134] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[135] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[136] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.